National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

NL-339 (r. 88/04) c

Canada

The University of Alberta

INTEGRATION OF CHINESE IN COMPUTER SYSTEMS

by

Hong Yuzhang

A thesis

submitted to the Faculty of Graduate Studies and Research

in partial fulfillment of the requirements for the degree

of Master of Science

Department of Computing Science

Edmonton, Alberta

Spring, 1989

National Library of Canada
Bibliothèque nationale du Canada

Canadian Theses Service
Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-53004-9

Canada

THE UNIVERSITY OF ALBERTA

RELEASE FORM

NAME OF AUTHOR: Hong Yuzhang

TITLE OF THESIS: Integration of Chinese in Computer Systems

DEGREE FOR WHICH THIS THESIS WAS PRESENTED: Master of Science

YEAR THIS DEGREE GRANTED: 1989

Permission is hereby granted to The University of Alberta Library to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only.

The author reserves other publication rights, and neither the thesis nor extensive extracts from it may be printed or otherwise reproduced without the author's written permission.

(Signed) Hong Yuzhang

Permanent Address:
HeiLongJiang Cadre College, QingAn,
HeiLongJiang, P.R. China.

Dated April 25, 1989

The University of Alberta

Faculty of Graduate Studies and Research

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research, for acceptance, a thesis entitled **Integration of Chinese in Computer Systems** submitted by **Hong Yuzhang** in partial fulfillment of the requirements for the degree of Master of Science.

W. A. Davis

Supervisor

Xiaobo Li

Ao Wai Hung

Jia-Huai You

Date: 24 Apr 89

To My People and My Family

ABSTRACT

This thesis discusses systematic approach to solve the Chinese computing problem. The central issue of Chinese computing, the compatibility problem, is studied based on the analysis of two typical Chinese software systems: CCDOS for the IBM-PC and 中文Talk (Chinese Talk in English) for the Macintosh. The advantage of graphics-based computers over text-based computers in Chinese computing is also investigated. A new concept, Character Channel, is introduced to systematically solve the Chinese/English compatibility problem.

Apart from Chinese system software, the thesis also covers many other important areas of Chinese computing, namely: computer representation of Chinese characters, Chinese word processing systems, electronic publishing, and machine translation. Chinese word processing systems, which best reflects the compatibility problem of Chinese computing, is studied in depth. Criteria for the evaluation of Chinese word processing systems are established that try to evaluate a Chinese word processing system in its entirety instead of treating it as a number of disjoint parts and then evaluating the parts separately. An improved Pinyin input method (i.e. Character Decomposition Pinyin) is also given together with the results of a number of experiments. A high quality character generating scheme is introduced as an addition to Chinese font libraries.

Acknowledgements

I am indebted for the advice and guidance of my supervisor, Dr. Wayne A. Davis. His knowledge and experience has guided me throughout this research and the preparation of the thesis.

I am grateful to the other members of my examining committee, Drs. Li Xiaobo, You Jiahuai, and J. So for their valuable comments and suggestions and time spent reading the thesis. I must thank the LiuHe Cadre College, BDRC (Black Dragon River Consortium) and CIDA (Canadian International Development Agency) for the financial support and help in a great many ways. Special thanks to Z. Tim for his great help especially in purchase of hardware, software and books. I am also grateful for the technical support provided by the Department of Computing Science. Special thanks to all my friends and teachers, especially Wang Yigong, Ping Zhenzhong, Yuan Xiaobu, Yuan Baoji, and Zhang Ge for the various form of help and encouragement that they have given to me.

I am also grateful to Kuo Hsiaohung, who contributed in a great many ways, especially in improving my English writing. Finally, I would like to thank Zhang Qiping, my wife, for her encouragement, patience and understanding.

Table of Contents

Chapter	Page
Chapter 1: INTRODUCTION	1
1.1 Motivation	1
1.2 Two Approaches to Chinese Computing	2
1.3 Compatibility - the Central Issue of Chinese Computing	2
1.4 Application software	3
Chapter 2: CHARACTERISTICS OF THE LANGUAGE	6
2.1 History	6
2.2 Comparisons with English	7
2.3 Some Features of Chinese Computer System	12
2.4 Conclusion	14
Chapter 3: COMPUTER REPRESENTATION OF CHINESE CHARACTERS	15
3.1 Chinese Information Interchange Codes	15
3.2 The Internal Machine Code Representations	20
3.3 Input Methods	22
3.4 Output Methods	32
3.5 Conclusion	36
Chapter 4: SYSTEM SOFTWARE	38
4.1 The Importance of Chinese/English Compatibility	38
4.2 CC-DOS	39
4.3 Zhong Wen Talk	48
4.4 Comparison of CC-DOS and Zhong Wen Talk	59
4.5 The Systematic Approach to Develop a Chinese System	61

4.6 A Graphics-Based Computer is Suitable for Chinese Processing 64
4.7 Other Systems 66
4.8 Conclusion 67
Chapter 5: CHINESE WORD PROCESSING SYSTEMS 69
5.1 A Brief Introduction to Three CWP 69
5.2 Criteria for Evaluation 72
5.2.1 Interpretation of Table III (Input Methods) 72
5.2.2 Pinyin Input Method 75
5.2.3 Interpretation of Table IV (Pinyin Input Methods) 77
5.2.4 Interpretation of Table V (Outputting Chinese) 80
5.2.5 Interpretation of Table VIII (Editing and Formatting) 87
5.3 Weight Assignment 90
5.4 Experimental Test of the Available CWP 95
5.5 Conclusion 102
Chapter 6: CHINESE ELECTRONIC PUBLISHING 103
6.1 Publishing Problem 103
6.2 Run-length code and radical composition methods 105
6.3 A Technique for generating high quality characters with high compression ratio 107
6.4 Character Generating Technique by Segments and
Complex Layout Generating Technique 114
6.5 The Whole Hua-Guang Publishing System 118
6.6 Other systems 122
6.7 Conclusion 124
Chapter 7: AUTOMATIC PROCESSING OF THE CHINESE LANGUAGE 125
7.1 Difficulties in Machine Translation 126

7.2 Available Computer Translation Systems 126
7.3 Conclusion 127
Chapter 8: SUMMARY OF RESULTS AND RECOMMENDATIONS 128
8.1 Character Decomposition Pinyin (CDPinyin) 128
8.2 Character Channel - a Solution to the Compatibility Problem 137
8.3 High Quality Character Generation as Compensatory Output Method 146
8.4 Conclusion 153
Chapter 9: CONCLUSIONS 155

References 158
Appendix A: Evaluating the Available CWP by Three Methods 161

List of Tables

Table		Page
I	Comparisons between Chinese and English	8
II	The Storage Capacities Corresponding to Dot-matrix and Number of Characters	36
III	Input Method Criteria	73
IV	Pinyin Input Method Criteria	78
V	Output Criteria	81
VI	The Usage-Frequency of Character Styles on Some Newspapers and Magazines	85
VII	The Character Styles Provided by Some CWPs and Chinese System	86
VIII	Editing/Formatting Criteria	88
IX	The Weight Assignment to Input Method Criteria	91
X	The Weight Assignment to Pinyin Input Criteria	92
XI	The Weight Assignment to Output Criteria	93
XII	The Weight Assignment to Editing/Formatting Criteria	95
XIII	The Results of Experiment 1 (News with 291 characters)	96
XIV	The Results of Experiment 2 (Poem with 176 characters)	98
XV	The Relation between Chinese Publishing Character Sizes and Their Dot Matrices	104
XVI	Some Examples of CDPinyin	129
XVII	Some Radicals with Different Names	133
XVIII	The Radicals with Simplified Names	133

List of Figures

Figure		Page
2.1	The Relation between the Shape, Sound, and Meaning of a Character	9
2.2	Overlapping two Components to Form a Character "Zhong"	10
3.1	Different Codes in a Computer System	16
3.2	GB2312-80 Distribution	18
3.3	Pinyin Codes with Different Length of Character Sequences	25
3.4	The Three Levels Indexing of Pinyin Code	26
3.5	The Head and Tail of a Chinese Character	28
3.6	Head/Tail Code Operation	30
3.7	The Structure of Head/Tail Code Indexing Lists	31
3.8	Three Ways to Represent Chinese Font	33
3.9	The 24*24 Dot-matrix for Chinese Character "Guang"	34
3.10	Display Chinese and English Characters on an Ordinary Screen	35
4.1	The Hierarchical Structure of MS-DOS	40
4.2	The Structure of CC-DOS	41
4.3	The Structure of Keyboard Management Module of CC-DOS	42
4.4	The Text Display Mode of MS-DOS	44
4.5	The Graphics Display Mode of MS-DOS	44
4.6	The Relation between VRAM and RRAM	46
4.7	The Procedure to Locate Cursor	47
4.8	The Original Macintosh Software System and ZhongWen Talk Architecture	49
4.9	The Script Manager Guides a Call from an Application to the Interface System	51

4.10	Some Relations Among the Modules of HIS	53
5.1	The Usage-frequency of Character Styles on People's Daily (Mainland China)	83
5.2	The Usage-frequency of Character Styles on Chinese Abroad (Taiwan)	84
5.3	A Piece from a Newspaper	97
5.4	A Classical Poem	97
5.5	The Comparison of Input Speed	101
6.1	Data Compression by Recording Black and White Segments	105
6.2	The Horizontal Stroke of Song Style Character	108
6.3	Two Irregular Strokes	109
6.4	The Contour Matrix, Flag Matrix and Final Dot Matrix for Output	111
6.5	Two-level-storage of Chinese Character Typefaces	114
6.6	Microprocessor F	117
6.7	The Hardware Configuration of Hua-Guang System	118
6.8	The Layout of the Example Annotation	121
7.1	The Language Automatic Processing System	125
8.1	The Character Channel for MS-DOS	139
8.2	The Character Channel for CC-DOS	143
8.3	The Character Channel for ChineseTalk	145
8.4	Some Strokes in Li-Style Stroke Library	149
8.5	The Character 中 ("center" in English) Formed by Stroke-Composition	150
8.6	Some Song-Style Characters with Different Sizes and Colors	151
8.7	Variations of Li-Style Characters by Scaling	152
8.8	The Artistic Characters Based on Li-Style Prototype	153

Chapter 1
INTRODUCTION

This thesis is concerned with computer processing in Chinese. The phrase "Chinese Language Computing" was virtually unheard of before the late nineteen sixties. A Chinese computing system is a computer system that the users can directly input Chinese into the system; process, store, and retrieve the information internally; and then output the result in Chinese.

Computer information processing techniques have been widely used in developed countries for many years because these countries invented and developed computer technology very early. In addition, these countries use alphabets which can be processed by computers easily. Whereas the Chinese language uses ideographic characters which have two important features: complicated shapes and an extremely large character set. There are several thousands of commonly used Chinese characters. Each and every character is ideographically different. It is these two features of Chinese characters that make Chinese computing a very difficult task. On the other hand, the development of Chinese computing in the past twenty years shows that Chinese computing challenges almost all areas of computer science from hardware to software, from main frame to auxiliary equipment, from system software to application software. In other words, Chinese computing presents significant problems for computer science. The work reported here represents an attempt to explore the unique features of Chinese computing to investigate a systematic way to the development of Chinese computing system, and to lay a foundation for further research in this area.

1.1. Motivation

For nearly twenty years, much research has been conducted on various aspects of Chinese computing, many Chinese computer systems have been produced and many papers on Chinese computing have been published. However, there have not been many papers that discuss Chinese computer systems in their entirety. Particularly, not many papers answer such questions as "What is the central issue of Chinese computing?" and "Is there any systematic approach to solve Chinese computing problem?" The thesis attempts to answer the questions based on an in-depth survey which includes Chinese system software, Chinese word processing systems, automatic processing of the Chinese language, Chinese electronic publishing systems, etc.

1.2. Two Approaches to Chinese Computing

There may be two approaches to Chinese computing: Patching approach and systematic approach. The Patching approach (named by the author) does not treat Chinese processing as an integrated part of computer system. It simply adds Chinese processing function onto an existing English system in some way. If some part of the system cannot work with Chinese properly, the part is isolated and a modified part is "patched" there. The idea behind the Patching approach is to consider Chinese processing as just a matter of input and output problem. The difficulty faced by this approach is that Chinese/English compatibility is not guaranteed and some good features of the English system may have to be sacrificed in order to accept Chinese. An example is CCDOS and its applications software which will be discussed in Chapter 4.

The systematic approach treats Chinese processing as an integrated part of computer system in order to obtain the best compatibility. To achieve this goal, some standard is necessary which should be followed by three groups: computer manufacturers, third party software developers, and users. The system should provide a solid foundation (i.e. hardware and operating system) for Chinese/English processing (based on the research for the thesis, a graphics-based computer is a good candidate). The advantage of this approach is the best compatibility so that software can be shared by both Chinese and English users. An example is 中文Talk and its applications software which will also be discussed in Chapter 4.

1.3. Compatibility - the Central Issue of Chinese Computing

Chinese/English compatibility is the central issue in the development of Chinese system software and Chinese computer systems. However, few papers on this subject have been published, although there are a few such papers on the situation from mainland China, they concentrate on the modification of existing English systems only [GW84]. This thesis intends to fill this gap.

Chinese/English compatibility is the central issue of Chinese computing because
(1) it affects all three aspects of Chinese computing, namely, input, processing, and output of Chinese/English mixed information;
(2) it affects all applications and it is so important that whichever application (e.g. word processing, database, network, etc.) is intending to use Chinese, the first problem to be dealt with is the Chinese/English compatibility issue.

Some people treat input and output as the central issue of Chinese computing. They argue

that so long as Chinese input and output is solved, a successful Chinese system is achieved. However, this is not true. For example, apart from I/O problem, determining internal Chinese code is also important which affects the entire performance of the system.

Compatibility problem has both theoretical and practical significance. Theoretically, successful solution to Chinese/English compatibility problem will directly lead to solution to multilingual system. Practically, it frees human from the burden of repeated software development. For example, 中文Talk results in most existing Macintosh software workable with Chinese. In this sense, KaiHin, a Chinese word processor written in C, and Mr. Zhang Ge's work, a Chinese system written in BASIC, become practically useless, because they used a Patching approach and failed to make use of the Macintosh full power.

Two typical Chinese system software for microcomputers are studied (i.e. CCDOS and 中文 Talk) due to the following considerations:

- CCDOS was developed in mainland China by **patching** MS-DOS (the original English system software) and it is a Chinese software system used extensively there. While 中文Talk was developed by the computer vendor itself (i.e. the Apple Computer, Inc. in U.S.) and it therefore forms an **integrated** part of the original system software. However, it has had little market in the mainland.
- CCDOS is based on the IBM-PC microcomputer which is a **text-based** computer, while 中文Talk is based on the Macintosh which is **graphics-based**.

Therefore, some important conclusions on Chinese system software can be drawn by analyzing and comparing these two systems, which are (1) graphics-based computers have obvious advantages over text-based ones in Chinese processing, (2) a standard is necessary to bring three groups (i.e. computer manufacturers, third party software developers, and users) together to achieve a compatible system. Based on the conclusions, a new concept, Character Channel, is introduced which attempts to solve the compatibility problem in a systematic way.

1.4. Application Software

Since the compatibility of a computer system is shown by its applications software such as Chinese word processor (CWP) and Chinese electronic publishing (CEP), the applications are also

discussed in the thesis.

Word processing systems are the most popular application software and involve many important aspects of Chinese computer systems such as Chinese input, Chinese output, and Chinese/English compatible processing. Unfortunately, not many Chinese computer scientists pay enough attention to them (this is especially true in the mainland). Some papers discuss input methods and Chinese keyboards [WEI85, LO85] but none (as far as can be ascertained) discuss Chinese word processing system in its entirety. In this thesis, very detailed criteria are established to evaluate Chinese word processing systems. In addition, the Pinyin input method (a phonetic Chinese input method by standard keyboard) is studied in more detail and an improved Pinyin input method (Character Decomposition Pinyin) is introduced.

Chinese electronic publishing (CEP) is another important application of Chinese computing due to its theoretical and practical significance. Theoretically, the most outstanding features of Chinese computing are reflected as its input and output aspects, and CEP shows the output feature to the full. Practically, replacing manual-typesetting by CEP can improve publishing efficiency greatly. This can also be seen by the fact: the first Chinese computing research project in China was established in August 1974 and among three research topics Chinese electronic publishing took the highest ranking (the other two were a Chinese information retrieval system and a Chinese communication system) [GUO85]. In this thesis, publishing problems are analyzed. The HuaGuang Publishing System, a successful CEP developed by Beijing University in mainland, is introduced in detail. The High Quality Chinese Character Generation for Casual Users (HQCG), an approach to generate high quality characters by using popular laser printing facilities, is also described.

The organization of the thesis is as follows: Chapter 2 discusses the characteristics of the Chinese language which is the foundation of the development of Chinese computer systems. Chapter 3 introduces computer representation of Chinese characters and the entire Chinese computer system is determined by the representation. Chapter 4 deals with system software, the most important part of the thesis, in which the Chinese/English compatibility problem is studied in detail. Besides CCDOS and 中文Talk, the implementation of Chinese system software in mini-computers and main frames is also introduced. Chapter 5 discusses Chinese word processing systems (CWP), another important part of the thesis, in which quantitative criteria for evaluating CWPs are established. Evaluation results of some typical CWP are given together with experimental test results. Chapter 6 introduces Chinese electronic publishing where HuaGuang system is introduced to show the complexity and difficulty faced by CEP as well as some

important techniques to deal with the problem. Chapter 7 deals with machine translation, another active area in Chinese computing. Chapter 8 summarizes the results and presents recommendations which include the CDPinyin (an improved Pinyin input method), the Character Channel (a suggested solution to the compatibility problem), and the High Quality Character Generation as compensation to font library. Finally, conclusions are given in Chapter 9.

Chapter 2
CHARACTERISTICS OF THE LANGUAGE

Chinese computing denotes the process of using a computer to process Chinese information. Since the Chinese language is very different from English it is necessary to study the characteristics of the language to set a basis for developing a Chinese computer system.

2.1. History

The Chinese language is one of the most ancient languages in the world with 5 to 6 thousand years of history. In the ancient Chinese characters which include the Jiaguwen (甲骨文), Jinwen (金文), Lishu (隶书) and Zhuanshu (篆书), the Jiaguwen (inscriptions on bones or tortoise shells) is one of the earliest script (C.16th - 17th century B.C.) and systematically completed. The most of basic features and character constructing rules of modern Chinese have already been formed in Jiaguwen [GUO85]. In 1880 the bones and tortoise shells with Jiaguwen on them were first found in AnYang, Henan Province. From that time until now about 3,000 Jiaguwen characters have been discovered.

The reason for the large Chinese character set is due to the following fact about pictophonetic characters in Chinese. About 3,000 years ago (C.16th century, the Shang Dynasty) Jiaguwen consisted of mainly ideographic characters (象形字 about 80 %) and pictophonetic characters (形声字 about 20 %). The former uses pictures to represent meaning and the latter uses one element to indicate meaning (meaning-symbol) and the other sound (sound-symbol).
For example, 牲 = 牛 + 生, where "牛" means "cattle" and "生" is the sound (pronounced as "Sheng"), namely, "牲" is something about livestock and pronounced as "Sheng".

Although the amount of pictophonetic characters started to increase at that time, they were not a major part of the Jiaguwen and the character form was not fixed.
For example, the meaning-symbol of the Jiaguwen character "牲" could be written as "牛" (ox), "羊" (sheep), "犬" (dog), etc.

Until 2,000 years ago Qin State unified China and then standardized Chinese writing, and the meaning-symbol of "牲" was fixed as "牛" (ox). However the other part of the pictophonetic characters, i.e. the sound-symbols, were never fixed.

Gradually, the sound-symbol of a pictophonetic character lost its function of indicating sound (the main reason is due to the various dialects in Chinese) and most meaning-symbols lost their function of indicating the category of things. For example, "骗" ("Pian" meaning to cheat) has a meaning-symbol "马" (horse) and obviously "cheat" has little to do with "horse".

Because pictophonetic characters were created in many different ways and for various purposes, many pictophonetic characters were created year by year. This is the main reason for the huge number of Chinese characters. In the book "Shuo Wen Jie Zi" (说文解字, "The Analysis of Characters" in English), Xu Shen collected 9353 characters (East Han Dynasty, 100 A.D.). Currently there are about 60,000 characters. So the number of characters increased by 50,000 which amounts to about 25 characters added each year.

After 1949, the founding year of the People's Republic of China, almost all added characters are simplified characters. Most of the classical characters and seldom-used characters are eliminated so that only 6,000 to 7,000 simplified characters are commonly used (although the classical characters are still kept in ancient books). In Taiwan and overseas Chinese communities classical characters are still commonly used, but there are no new characters added to them. Nowadays, more and more new phrases are added into the Chinese language but not new characters. In short, the large Chinese character set resulted from the rapid increase of pictophonetic characters in the Chinese language.

2.2. Comparisons with English

The fundamental problems of the Chinese language processing by computers are:

To input the massive number of ideographic Chinese characters into a computer system is not easy. Even if one solves the problem, it is important to be sure that the correct characters were entered. Then come the questions of: how to display and output the Chinese characters, how to represent the Chinese characters in binary, and how to handle internal processing, manipulating, storing, retrieving, and communication of the information [YU88].

In order to have a better understanding of these problems, comparisons between Chinese and English are listed in Table I followed by an explanation.

Table I. Comparisons between Chinese and English

Characteristics	Chinese	English
1. Representation	ideographic	spelling
2. Character shape	complicated	simple
3. Character set	huge	small
4. Character style	calligraphic and artistic, classical form and simplified form	mainly artistic, upper case and lower case letters
5. Writing direction	horizontal and vertical, right-to-left heading in some cases	mainly from left to right
6. Homonyms	many	not many
7. The number of monosyllabic words	many	very few
8. Indexing and sequencing	difficult, by radical, Pinyin, and usage-frequency	easy, by alphabet

The following are explanations of Table I:

1. Representation

Chinese characters are ideographic in nature and have a two dimensional structure. Whereas English is a language with each word composed of letters taken from an alphabet.

The common nature of human languages is to represent meaning by the sound and shape of symbols. The relationship between shape, sound and meaning is shown in Fig. 2.1. To know the

meaning of a character (or word in English) people can either "see the shape and know the sound, then know the meaning by sound" or "know the meaning by shape" directly.

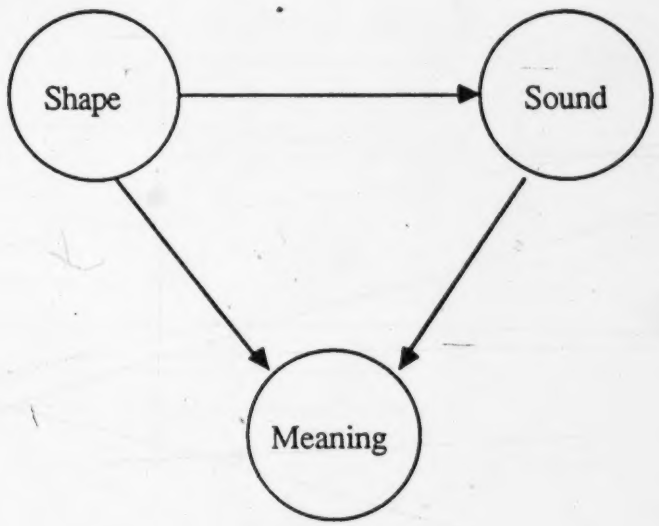


Fig. 2.1 The Relationship between the Shape, Sound, and Meaning of a Character

A spelling language emphasizes the relation between sound and meaning, that is, people know a word's pronunciation by looking at its shape and then understand the meaning from the pronunciation. While an ideographic language emphasizes the relation between shape and meaning in the sense that very often people do not know the pronunciation of a character, but know its meaning by looking at the shape. In Chinese, most characters cannot be pronounced by their shape but the meaning of many characters can be understood just by looking at the shape.

For example, 女 ("nu" meaning woman) developed from the ancient character " 𡚁 ", and its meaning is known by looking at the shape which looks like a woman.

烯 ("Xi" meaning alkaline in chemistry) is unknown to most people. However they know it must be something combustible because the meaning symbol (left part) is "火" which means "fire".

A Chinese character consists of components; however, it is not constructed by a one dimensional sequence of components (note: an English word is an one dimensional sequence of

letters). Instead, a Chinese character is formed by overlapping components in a two-dimensional space (Fig. 2.2). This implies that a Chinese computer must be a graphics-based machine in order to fully represent Chinese characters. This will be further explored in Chapter 4.

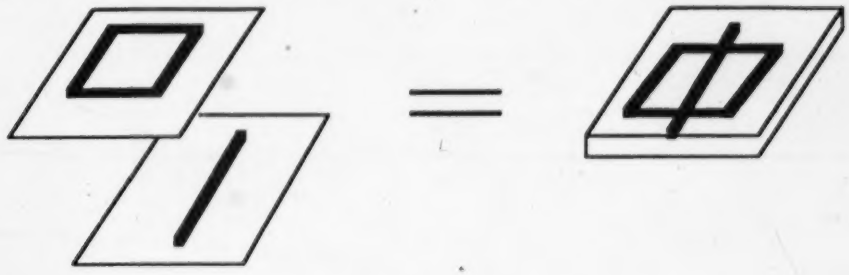


Fig. 2.2 Overlapping Two Components to Form a Character "中" ("Zhong" meaning center) [ZHANG84]

2. Character shape

It is obvious that Chinese characters have a complicated shape due to their pictorial nature.

3. Character set

The Chinese character set is extremely large, for example, GB-2321-80 (The Government Standard) consists of 6,373 characters. Whereas there are only 26 letters in the English alphabet. It is the complicated shape and the huge character set that make Chinese processing a very difficult task.

4. Character style

Multi-style is another feature of Chinese characters, which consists of calligraphic style and artistic style. Calligraphy is not only a kind of art, but also widely used in ordinary writing and publishing. Although both calligraphic and artistic styles exist in English, only artistic styles (e.g. italic or bold style) are widely used in ordinary writing and publishing. It is this feature of Chinese characters that makes Chinese output a difficult task.

5. Writing direction

Usually, English is written horizontally from left to right, while Chinese can be written vertically without any inconvenience because Chinese consists of square characters (e.g. in Taiwan

and overseas Chinese community vertical-writing is more popular than horizontal and many headings and subheadings of publications are printed horizontally from right to left). This feature adds another level of complexity to the output of Chinese.

6. Homonyms

A large number of homonyms shows another important feature of Chinese that is related to the fact that there are thousands of characters (e.g. 42,174 characters in KangXi Dictionary and 33,000 in Taiwan's CCCII Interchange Code), however, there are only 423 sounds (or 1,351 sounds with tones added). Therefore, many characters have the same pronunciation. This has a negative effect on Chinese input and indexing.

7. The number of monosyllabic words

In English there are very few monosyllabic words (e.g. "a" and "I"), whereas, in Chinese most characters are monosyllabic words (单音词). For example, among 8500 characters in the New China Dictionary 7922 characters are monosyllabic ones (93.2 %) and only 578 characters are not (6.8 %), e.g. 葡 and 萄 are not monosyllabic words (although they are monosyllabic characters) and they must be used together to form a word, "葡萄" ("pu tao" meaning grape).

The massive number of monosyllabic words in Chinese has a great advantage over English in the sense that Chinese is very powerful to form phrases (words). That is why there are large numbers of two-character phrases and four-character idioms in Chinese. For example, 兄 ("xung" meaning elder brother) and 弟 ("di" meaning younger brother) can form "兄弟", "弟兄", "弟弟", etc. This feature of Chinese provides the basic way to solve the homonym problem in Chinese input, which will be discussed in Chapter 5.

8. Indexing and sequencing

Indexing and sequencing (sorting) of English words is very simple (i.e. alphabetic order), while in Chinese there is no simple way to sort characters. First, the massive number of homonyms make indexing by sound impossible. Second, Chinese contains hundreds of radicals and one radical can be composed of others (no exclusivity), and a word can be formed in several ways (no uniqueness). It has no standard collating sequence. Therefore, it has no unique classification by radicals. A Chinese character (AB) composed of radicals A and B may appear to fall under the radical A, but sometimes under the radical B. Therefore indexing by shape only is

not practical.

As a result, Chinese characters are usually sorted by a combination of several ways, for example, in GB-2312-80 (The Government Standard) characters are sorted by a combination of radicals, sound (Pinyin) and usage-frequency.

2.3. Some Features of Chinese Computing Systems

Based on an analysis of the characteristics of the Chinese language, the features of Chinese computing systems can be better understood. On one hand, Chinese and English are treated equally by computers in the sense that both of them are represented by a sequence of bits, i.e. 0's and 1's. So, a Chinese computing system could (and should) make full use of the functionality of English computing systems and hence Chinese/English compatibility is an important feature for all Chinese computer systems (see Chapter 4). On the other hand, due to the obvious differences between Chinese and English, Chinese computing systems have unique features which challenge almost all areas of computing science:

(1) Computer System Architecture

The conventional von Neuman computer is a sequential-execution one which is suitable to alphabet-sequencing of spelling languages. A small keyboard and a simple character generator with a fixed character set are enough for English input and output.

When processing Chinese, a computer has to work in graphics mode in order to output Chinese characters. It is not difficult to imagine that if computer were invented by Chinese people, the computer would be an image computer or graphics computer. It would provide an instruction set of graphics operation (e.g. overlapping in Fig.2.2) and a set of image data structures (or graphics data structures) based on icons. In addition, since a Chinese character is composed by overlapping components the CPU may be a parallel array of processors.

In Chapter 4, it is shown that a graphics-based computer (like the Macintosh) has obvious advantages over a conventional text-based computer (like the IBM-PC) in Chinese processing.

(2) System software

The available system software, ranging from mainframes and minicomputers (e.g. UNIX) to microcomputers (e.g. MS-DOS), are written in English, and cannot work with Chinese (especially its I/O subsystems). Therefore, how to develop a Chinese/English compatible system software so that all application software above it can use both languages is the central issue of Chinese

computing.

In fact, if a Chinese/English compatible system can be successfully designed, a multilingual system may be achieved. The key issue of compatibility is to design the language-interfaces in a systematic way so that all languages in the world (or at least the major languages such as Chinese and Arabic) can have easy access to the system software and the application software above it. This is undoubtedly a significant research area (see Chapter 4).

(3) Artificial Intelligence (AI), Pattern Recognition and Image Processing

AI is widely used in Chinese Language Automatic Processing, especially machine translation, which is introduced in Chapter 7.

Since Chinese input is a bottle-neck of Chinese computing, various techniques are invoked in this area which includes: AI, knowledge base, pattern recognition, etc. Basically, there are two categories of Chinese input methods: keyboard or pattern recognition.

To allow ordinary people to input Chinese by keyboard, a computer system has to be tolerant of incorrect or partial Pinyin (due to different dialects of Chinese), different input-sequences of radicals or components, phrase-input, etc. As a result AI and knowledge base techniques are widely used.

Non-keyboard input techniques include speech recognition and character recognition. It is obvious that a Chinese computer system with speech input would be accepted by all Chinese people. Unfortunately speech recognition is one of the most difficult tasks of computer application. It is especially true for Chinese speech because of many different dialects of Chinese (which are so different that many Chinese people cannot understand each other if they use their own dialects). It also challenges AI because the speech input system must be an adaptive one with learning ability.

For a long time, Chinese computer scientists have paid great attention to character recognition because:

- the pictorial nature of Chinese character implies that keyboard-input is not easy;
- Chinese people do not have the tradition of using typewriters and so they are not so comfortable with a keyboard;
- character recognition is especially suitable for inputting Chinese in massive manner.

Now, a Chinese Character Recognition (CCR) system for printed-character recognition is available, but handprinted CCR is far from being practical [YU88]. Furthermore, even if a commercial handprinted CCR is available ordinary people may not accept it because they still have

to write very neatly which implies slow speed.

(4) Output techniques and electronic publishing

The pictorial nature of Chinese characters also challenges output techniques especially for electronic publishing. Just like English letters, Chinese characters have to be generated by either fetching the prestored font or creating them on-line. Considering the massive number of characters with many styles and sizes, neither is an easy task. For example, for characters with a 48*48 dot matrix, the set of 6763 characters (GB-2312) will take 1.9 MB for just one font. That is why image compression/decompression has been an active research area for a long time.

(5) Networks and distributed systems

Computer networks and distributed systems are necessary for Chinese processing. The questions raised by Chinese computing are:

- How to modify the operating system and protocols of a network to make it accept Chinese?
- How to distribute stations, font libraries, data bases and software libraries to keep a relatively balanced load on a network with high efficiency?
- What network topology is suitable for Chinese processing?

Due to the fact that networks are not currently popular in China and very often people just select a popular network such as Ethernet and use it (they seldom consider whether the network meets the requirement of Chinese processing), those questions have neither been answered nor often even asked.

2.4. Conclusion

In summary, Chinese computing challenges many areas of computer science due to the characteristics of the language. Because of space limitations of the thesis, it will not cover all areas mentioned above. Even so, an important idea results from the thesis, that is, Chinese computing should be an integrated part of computer systems, or put it in another way, a computer system should have language compatibility with its hardware and software accepting both spelling and ideographic languages.

Chapter 3
COMPUTER REPRESENTATION OF CHNESE CHARACTERS

The symbols used in any natural language are represented as certain codes (i.e. binary patterns) in computer systems. In order to facilitate communication among different components (hardware and software) within a computer system as well as between different computer systems, the standardization of the code is inevitable. These standards are termed as a Code for Information Interchange (abbreviated as CII). They are normally sanctioned by the standards organization of the respective countries as well as by the International Standards Organization (ISO).

For Latin based national languages, there are several standards available today. One example is the American Standard Code for Information Interchange, the 7 bit ASCII code.

With regard to Oriental Languages, such as Chinese, Japanese, and Korean, they are different national languages. Japan's Industrial Standard (JIS) is the earliest among all oriental countries to come out with a standard of CII. JIS C 6226 was defined in 1978 and was amended in 1983 to have a total of 6,877 characters, with 2,965 in the First Level, 3,399 in the Second Level, and 524 non-Kanji characters [YU88]. Similarly, to process Chinese by computer, Chinese character encoding is the first problem to be solved. From the view of encoding Chinese, a Chinese information system is actually a conversion system which performs conversion between different Chinese codes (Fig.3.1). Chinese codes consist of: Chinese interchange code, internal code, input code, address code, font code, etc. These codes will be introduced below.

3.1. Chinese Information Interchange Codes

Today, there are many standards of CII for Chinese. Different CIIs emerged to meet the needs of national language requirements of different countries and localities. The PRC's standard GB2312-80 was published in 1981. This Chinese graphic character set code for information interchange defines a total of 6,763 Hanzi characters (with 3,755 characters in the First Level, and 3,008 characters in the Second Level) and 682 non-Hanzi characters. In Singapore, the Chinese character set used is the same as the PRC with the exception of a few new Chinese characters created for use in Singapore.

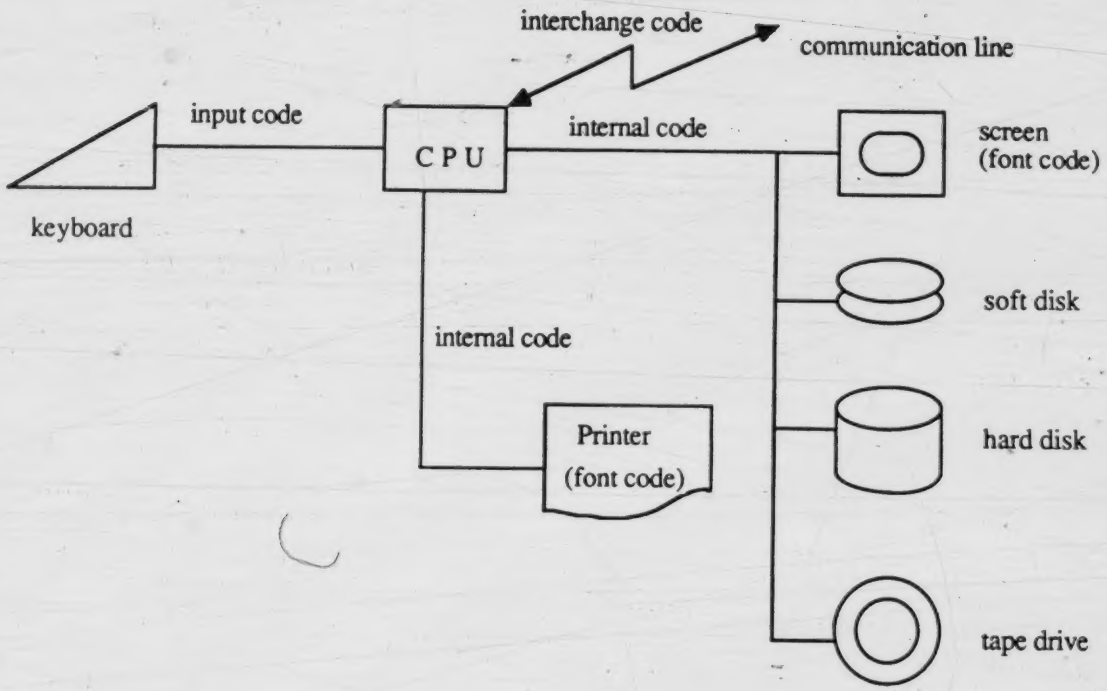


Fig. 3.1 Different codes in a Chinese computer system

In Taiwan, the Taiwan National Library published a 3 byte Chinese Character Code for Information Interchange (CCCII) in 1980 in an attempt to accommodate more than 50,000 Chinese characters with traditional as well as variations and simplified forms. This 94 planes of 94*94 code positions has not been accepted for practical reasons by computer manufactures with the exception of the Research Library Group in Stanford University for the use of library applications. Again, in Taiwan, in March of 1986, the government adopted the code of information interchange which has been widely used and implemented by the Chinese computer manufactures and vendors in Taiwan as a standard. This standard, The General Code Standard, has two code pages with 13,051 Hanzi characters. The first code page contains 4,808 frequently used Chinese characters and the second code page contains the remaining Chinese characters.

When designing an interchange code for Chinese, some unique problems have to be carefully dealt with.

Problem 1: Compatibility

Since most Chinese systems developed today are based on available computer systems, a Chinese interchange code should be compatible with ASCII code, i.e. ASCII code should be a

subset of the Chinese interchange code. The advantage of a compatible system is that Chinese people can make full use of the available English software and the software developed in China can be shared by people in the rest of the world.

Since two bytes are needed to represent a Chinese character while the ASCII code needs only one byte, the second byte of a two-byte code is used as an ASCII code in PRC's GB 2312-80 Code Table.

Problem 2: Practicality, universality and encoding-efficiency of Chinese character selection.

The features of Chinese characters include the very large character set, complicated shapes, and numerous homonyms. Therefore, for any Chinese information system the first question to be answered is how many and which characters should be chosen and included in the CII. If the information interchange code contains too many characters, then the hardware and software overhead will not be affordable. If it does not contain enough characters, the basic requirement by ordinary people cannot be met. This is the so called practicality problem.

The usage of Chinese characters is complicated. There exist big differences among the characters used in PRC, Taiwan, Hong Kong and overseas Chinese communities. Also there are big differences among the Chinese characters used by different professions and industries. This is the so called universality problem.

According to the above considerations, the Chinese interchange code GB2312-80 (GB means Government standard) was published by the General Bureau of National Standards of the PRC in 1981. GB2312-80 uses the ASCII code set as its basic set and uses two bytes for the Chinese interchange code. In the following part of this section GB2312-80 is introduced in detail because it is a widely accepted CII in both China and the world.

(1) Section and Position

In the GB code table, the first byte of a Chinese interchange code is for "section" and the second byte is for "position". There are a total of 94 sections with 94 positions for each of them.

The non-Chinese graphics-character area occupies Sections 1 to 15, i.e. the first byte is hex(21) to hex(2F) which contains 202 general purpose symbols (e.g. space, punctuation marks, operation symbols, etc.), 60 ordinal number symbols, 22 number symbols, 52 Latin characters, 169 kana (the Japanese syllabary), 48 Greek characters, 66 Russian Characters, 26 Pinyin Characters, 36 Chinese phonetic symbols.

The area from Sections 16 to 94, i.e. for the first byte of hex(30) to hex(7F) is the Chinese character area which contains 6763 Chinese characters (divided into 2 levels). There are 3755 characters in the First Level ranging from Sections 16 to 55, i.e. hex(30) to hex(57). These

characters are the most frequently used characters sorted by their Pinyin spelling. There are 3008 characters in the Second Level ranging from Sections 56 to 87, i.e. hex(58) to hex(77). These characters are less frequently used and are sorted by strokes or components. Sections 88 to 94 are empty (Fig. 3.2) [ZHANG84].

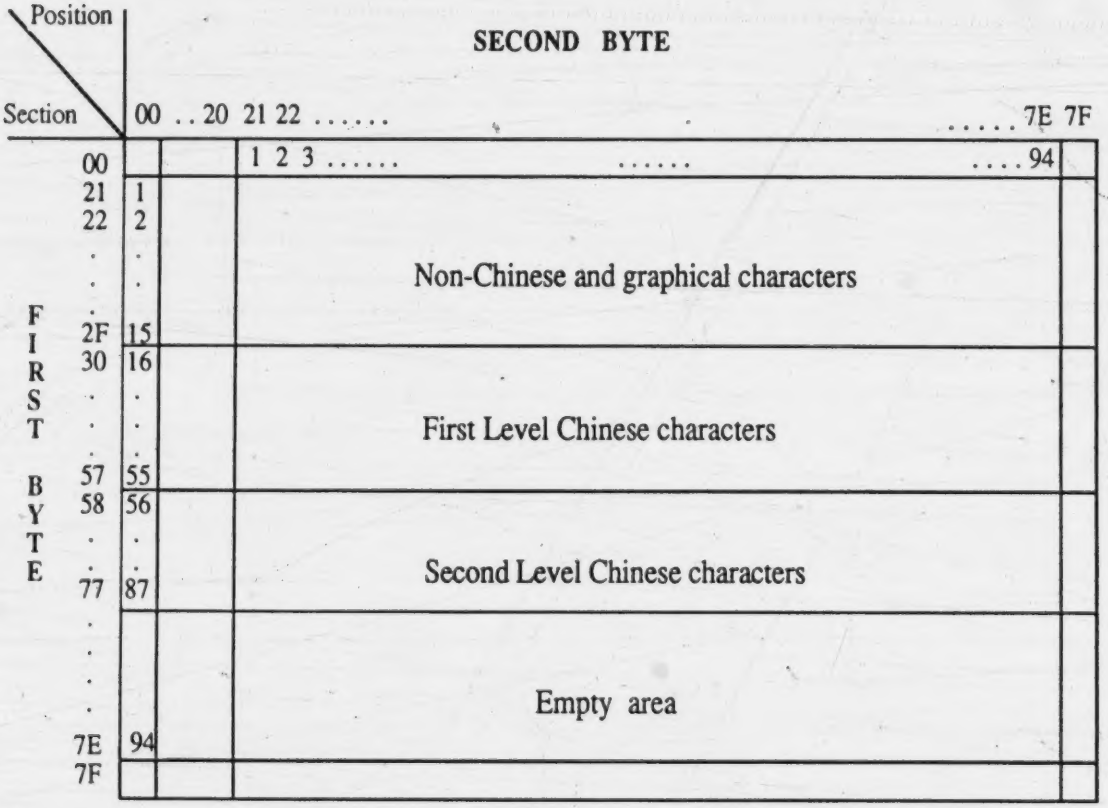


Fig. 3.2 GB2312-80 Distribution [ZHANG84]

(2) Selecting and Sorting Chinese Characters

The 6763 characters of GB2312-80 were chosen based on the "Popular Chinese Character Set for Publishing Purpose" (6196 characters, released by the Cultural Ministry of PRC and Chinese Character Reformation Committee in 1965), plus more than 500 characters which are commonly used as scientific and technological terms, person's names and the names of places. For most

Chinese information systems, so long as they support 6763 characters in GB 2312-80, the basic requirement can be satisfied. This is significant for reducing the time and space cost of a Chinese information system, shortening the code-length of characters and improving the efficiency of character encoding.

According to statistical analysis of usage-frequency of Chinese characters in 1975 in China, the 6763 characters of GB2312-80 satisfy 99.99 % of the text. The usage-frequency of these characters are very different, that is, 3000 to 4000 frequently used characters satisfy 99.9 % of the text. Therefore, a Chinese system with 3000 to 4000 such characters can meet the basic requirement. To ease the usage of characters and to manufacture Chinese computer systems of different levels, it is necessary to classify these 3000 to 4000 characters as First Level characters. Finally, 3755 characters are chosen as First Level characters based on usage-frequency, ability of phrase-construction, and practical application. The remaining 3008 are Second Level characters.

The characters in Level 1 and 2 are sorted in different ways. Since the First Level characters are the most frequently used characters with their pronunciation known to ordinary people, they are sorted by Pinyin (Chinese phonetic alphabet). For a character with multi-pronunciation the most popular pronunciation is taken. Homonyms with the same tone (note that there are 4 tones in the Chinese language) are sorted according to their first stroke and if they have the same first stroke, then the second stroke is compared, and so on.

Since the characters in the Second Level are less frequently used, it is not easy to pronounce most of them, and they are sorted according to radicals (in the way similar to a Chinese dictionary). Radicals are sorted according to the number of strokes. The characters with the same radicals are also sorted in this way. If characters have the same number of strokes, they are sorted according to stroke order (i.e.— | ノ 乀).

(3) Application of GB2312-80.

GB2312-80 is widely used in information interchange and hardware/software designs. The following are some examples:

Most Chinese font libraries are designed based on GB2312-80. The character selection for large keyboards, input code conversion, and character address code for output devices are based on GB2312-80. Besides, most high level languages with a Chinese processing capability, Chinese Database systems, and Chinese information retrieval systems are also based on GB2312-80.

In addition, since GB2312-80 is the fundamental standard of Chinese information processing, many other standards are closely related to it. For example, the standard for Chinese fonts (bit

map), the standard format of tapes and disks, and the standard for input/output devices are all based on GB2312-80.

3.2 The Internal Machine Code Representation

For a variety of reasons, the actual binary representation used in a computer system may not be exactly the same as the codes defined by a CII. For example, all large IBM systems use EBCDIC and not ASCII. Therefore, for Chinese computers, one should not confuse CII with the internal code. Standardization of Chinese CII is essential. A set of commonly acceptable internal machine codes may be desirable, however, standardization may have the shortcoming of putting unnecessary restrictions on future machine architectures.

The criteria for designing an internal code are:

- the code-length should be as short as possible, while the number of Chinese characters represented by the code should be as big as possible.
- the code values are continuous and it is easy to manipulate the code.
- the code should be compatible with the standard interchange code (i.e. there should be a very simple relation between the code and the standard interchange code).
- the code should be easy to be accepted as a character string data type for high level languages.

Two popular approaches (in PRC and Taiwan) are introduced below followed by the detailed description of the double-byte and four-byte internal representation.

(1) Both high order bits on:

This is used by many computer vendors to allow the coexistence of a double-byte code with 7-bit ASCII codes. It distinguishes a double-byte character from a normal single-byte character by turning the 8th bit of both bytes in the GB2312-80 codes to one, i.e. from the code space of 2121(hex) to 7E7E (hex) to the code space of A1A1 (hex) to FEFE (hex). This internal code representation is currently the most popular in China [YU88].

(2) The Big-5 code:

This is the most popular double-byte internal code representation among all Chinese computer manufacturers in Taiwan. It was published by the Institute of Information Industry of Taiwan in May 1984. The range of the First Byte is from A1(hex) to FE (hex) and the Second Byte is from

40 (hex) to 7E (hex) and A1 (hex) to FE (hex). The total number of code spaces is 94*(63+94) = 14,758. The assignment of the characters does not follow the sequence of the General Code Standard of Taiwan [YU88].

The following example shows how a double-byte internal code is formed based on GB2312-80.

For the character "[illegible]"

	Binary	Hex
GB2312-80:	00110000 00100001	3021
Internal code:	10110000 10100001	B0A1

The advantage of a double-byte internal code is that it is the shortest Chinese internal code, so it takes minimum storage space and improves transmission speed. However, it cannot be used in the environment where the 8th bit is used for other purposes (e.g. as a parity bit in communication).

The Four-byte internal code is also based on GB2312-80 and it is formed by decomposing and composing the GB2312-80 code and adding a Chinese flag byte.

Assume GB2312-80 is represented as

$$0\ B_{17}\ B_{16}\ B_{15}\ B_{14}\ B_{13}\ B_{12}\ B_{11}$$
$$0\ B_{27}\ B_{26}\ B_{25}\ B_{24}\ B_{23}\ B_{22}\ B_{21}$$

where B_{ij} (i = 1,2; j = 1,2, ..., 7) represents a bit in a GB2312-80 code.

The procedure of generating a four-byte internal code is the following:

(a) Decomposition.

Divide the two-byte GB code into a three-byte code in the following way,

$0\ 0\ 0\ 0\ B_{17}\ B_{16}\ B_{15}\ B_{14}$ (First byte)
$0\ 0\ 0\ B_{13}\ B_{12}\ B_{11}\ B_{27}\ B_{26}$ (Second byte)
$0\ 0\ 0\ B_{25}\ B_{24}\ B_{23}\ B_{22}\ B_{21}$ (Third byte)

(b) ASCII mapping

Map the three-byte code to 0 ~ 9 and A ~ Z of ASCII code, that is,

if a byte is less than 09 then add 30 (hex) to the byte,

if a byte is greater than 09 then add 37 (hex) to the byte.

For example, for the Chinese character "[illegible]"

	Binary	Hex
GB2312-80:	00110000 00100001	3021
First byte:	00000110	06
Second byte:	00000001	01
Third byte:	00000001	01

Since all three bytes are less than 09 in the example, 30 (hex) is added to all of them, results:

	ASCII code	Meaning
First byte:	00110110	6
Second byte:	00110001	1
Third byte:	00110001	1

The final four-byte code: 00001111 00110110 00110001 00110001
where the first byte is the flag byte (Chinese character indicator).

The advantage of four-byte code is that it has much wider application and good adoptability. However, it takes more storage space and slows down transmission speed. Besides, displaying and editing are difficult compared with a two-byte internal code.

For most Chinese systems in PRC, the two-byte internal code is strongly recommended and only in special cases, is the four-byte internal code used as an alternative.

3.3 Input Method

The Chinese input method (also called input code) is the means of entering Chinese characters into a computer system and which is located at the human-computer interface. More specifically,

input code is the interface between Chinese characters and keyboard. It maps tens of thousands of characters onto tens of keys of the keyboard. Since Chinese characters are input into computers mainly by a keyboard, input methods (or input code) should be very user-friendly.

To input a Latin alphabet (like English) is very easy in that the input code can be generated by pushing different keys on a keyboard (note that each Latin letter and symbol is represented by a distinct key on the keyboard).
Therefore,

{Input code} = {Internal code}

Input of Chinese characters could be the same way, that is,

{Chinese Input code} = {Chinese Internal code}

This is the way the large keyboard input method works. Because each key of the keyboard corresponds to a Chinese character and there are thousands of such characters, the keyboard has at least 2,000 keys. Obviously, only well-trained professionals can use it effectively.

Except the large keyboard input method, any Chinese processing system must generate a Chinese input code (which is different from the internal code) by input devices (a keyboard is most commonly used). As previously mentioned, such input methods must be very user-friendly. The most commonly used criteria for input are:

- a simple (and easy to remember) encoding rule,
- easy to operate,
- able to represent a large character set,
- short code-length,
- high input speed,
- low rate of duplicate-codes,

Many computer scientists think that the input method is the "bottleneck" of Chinese computing in the sense that only when Chinese characters can be entered into computers easily, can computers be widely used by ordinary people. For years, considerable effort has been put into this research area and four to five hundred input methods using a standard English keyboard have been developed. These may be classified as three categories: calculating-encoding, non-duplicate-code encoding, and duplicate-code encoding [GW84].

(1) Calculating-encoding

For the calculating-encoding input method the conversion from the input code to internal code is carried out by calculation. The codes have equal length. There is no disk access and no need to look up a code table. Therefore, it takes less memory apace and the speed of conversion is high. However, since codes are sorted according to their positions in the font library, there is no rule to help remember them.

(a) GB code:

For the GB code input method, users type the GB2312-80 code directly (i.e. four hexadecimal numbers) and the system converts the GB code to the internal code by changing the 8th bits of both bytes to one.

(b) Section/position code:

This is a meta-GB code in the sense that there are 94 sections with 94 positions for each section in the GB code table and so the section/position numbers (decimal numbers) have a very simple relation with the GB code. Users type four decimal digits (two for the section number and two for the position number) and computer converts the code to the internal code in the following way:

- convert decimal section/position code to the hexadecimal code,
- convert the above code to the GB code by adding 20 (hex) to it,
- change the 8th bits of both bytes of GB code to one (which is the internal code).

(2) Non-duplicate-code encoding

A Duplicate-code means that an input code corresponds to more than one character (e.g. a Pinyin input code can represent up to 200 homonyms). A typical non-duplicate code is the telegram code which uses four decimal digits to represent a Chinese character, so it is one to one mapping. Although the telegram code has been used in China for a long time, only professionals can remember the code for the thousands of Chinese characters.

(3) Duplicate-code encoding

Most duplicate-code encodings are user-friendly due to the fact that the simpler an input-rule is, the more user-friendly the input method will be. However, it is almost impossible for a very simple rule to distinguish all Chinese characters (i.e. a duplicate code is inevitable). Therefore human-computer interaction is used to solve the duplication problem.

The central point for the conversion of a duplicate-code encoding is a hierarchical search. And the data structure of the conversion table is the key to understanding the conversion.

Two typical such encoding methods are the Pinyin code (which inputs characters by sound) and Head/Tail code (which inputs characters by shape).

3.3.1 Pinyin Code

The Pinyin code is a very user-friendly input encoding method. When entering a character, a user types Pinyin letters on the keyboard and all the characters corresponding to this Pinyin sequence will be displayed on a hint line at the bottom of the screen (usually 5 to 10 characters a line). If the character wanted is on the line, the user just selects the number key corresponding to the position of the character to choose it. If the character does not appear on the hint line, the next group of characters can be shown on the hint line by pushing a defined key or mouse button.

The Pinyin code has to deal with some problems in some cases. For example, the length of Pinyin sequences are not the same, which may range from 1 to 6 letters (Fig. 3.3). Also, some users may not be able to remember the correct Pinyin sequence for a character or may remember only part of the sequence. Usually a Pinyin encoding method can be tolerant to incorrect or to some extent a partial Pinyin sequence.

A	阿
AI	爱
BAI	百
WANG	王
ZHONG	中
ZHUANG	装

Fig. 3.3 Pinyin codes with different length of character sequences.

Because Pinyin method is a duplicate-code encoding, human-computer interaction (by a hint line) is adopted to help a user choose characters. As a result, it becomes a very simple and easy-to-use input method (probably the easiest and simplest in the author's opinion) which is so easy to use that any student from high school (even elementary school) can enter characters by Pinyin without training.

The indexing table of the Pinyin code is based on GB2312-80 (note that the First Level characters are arranged in Pinyin order). As long as the boundary point of two duplicate code

groups can be found, the hint line can be built (i.e. one duplicate-code group per hint line assuming the line is long enough).

Pinyin code indexing consists of 2 levels: First Level indexing and second level indexing (Fig. 3.4). The First Level is a Pinyin code table with each entry containing a pointer to a GB-Pinyin-duplicate-code-group and each entry is 4 to 6 bytes long (corresponding to the Pinyin sequence with 4 to 6 letters). For Pinyin codes shorter than 4 letters, the space-bar is used to indicate the end of the code. The second level is a GB code table with two bytes for each entry which also holds the pointer to the GB-Pinyin-duplicate-code-group, that is, the entry holds the GB code corresponding to the Pinyin code of the First Level table. In other words, both First Level and second level lists contain pointers to the duplicate-code-group, but the pointers in the First Level list is Pinyin code and the pointers in the second level list is GB code.

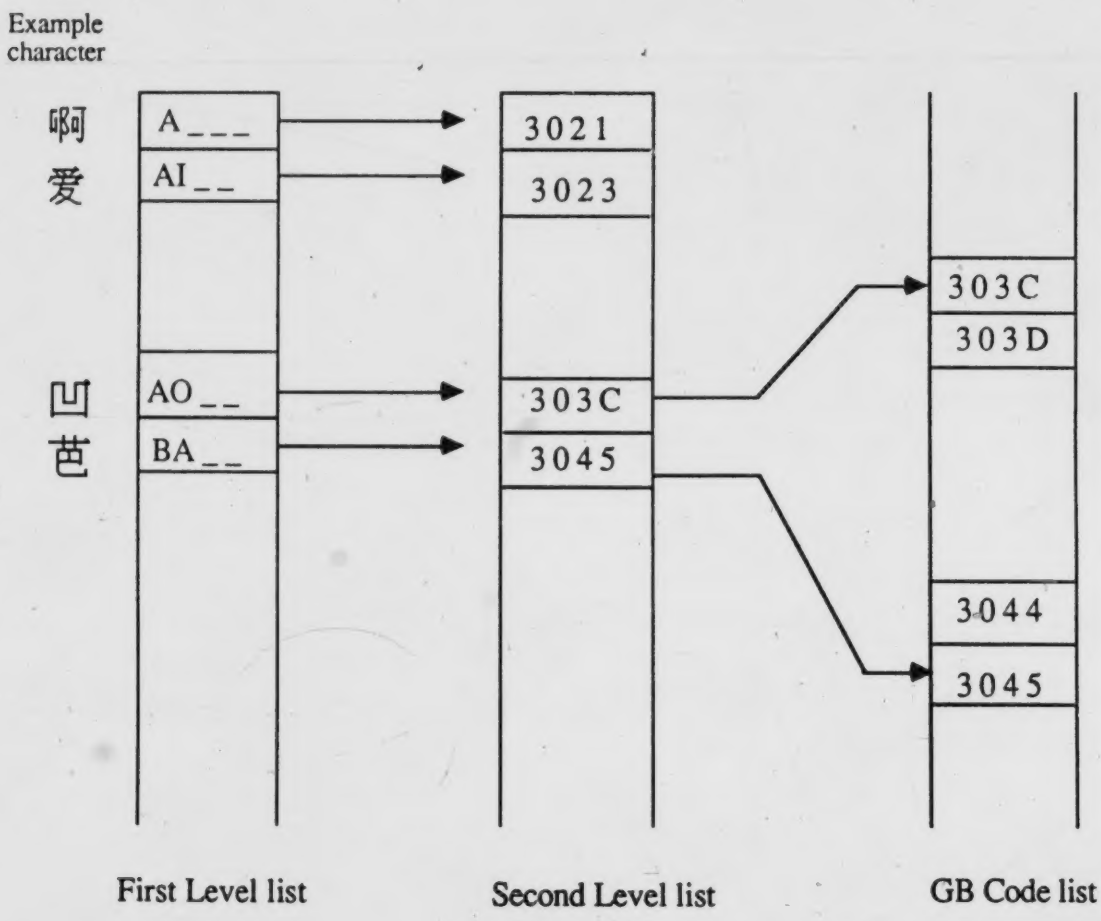


Fig. 3.4 The three levels indexing of Pinyin code [GW84]

Fig. 3.4 shows its structure. By sequential or binary searching on the first level table the pointer on the second level table can be found. Then the difference between two adjacent pointers on the second level table indicates the number of the duplicate-code (i.e. the number of homonyms for a given Pinyin code). Then these homonyms can be fetched and displayed on the hint line for users to choose. For example, a user types Pinyin code "AO" with its pointer pointing to an entry in the second level table (i.e. 303C which is the first GB code in the duplicate-code group for "AO"). The difference between two adjacent entries, i.e. 3031 and 3045 indicates the number of the duplicate codes in the group.
In this case,

$$3045 - 303C = 9$$

Namely, there are 9 duplicate codes in the group corresponding to 9 homonyms. They are:

303C	凹
303D	敖
303E	熬
303F	翱
3040	袄
3041	傲
3042	奥
3043	懊
3044	澳

The homonyms will be shown on the hint line when a user types the Pinyin code "AO".

It can be seen that although Pinyin has a high duplicate-code rate, it is very intuitive and users can search the GB code sequentially. So, it turns out to be an important and popular input method.

3.3.2 Head/Tail Code

A restriction of Pinyin code is that it can only be used by those who know the standard Chinese

pronunciation (i.e. mandarin Chinese). For the Chinese people speaking local accents, Pinyin is not convenient. Also, the nature of Chinese characters is more pictorial than phonetic. Very often, the sounds of some characters are unknown to many, but they can write (or read) the characters without difficulty. Therefore hundreds of shape-based input codes have been developed. The Head/Tail input code is a typical one.

In the Head/Tail code, 36 most commonly used radicals are arranged on a keyboard based on the ideographic nature of Chinese characters and the numerical keys from 1 to 5 are for the interaction with users. The Head/Tail code can be defined as follows:

The input code consists of 3 parts: the Head code, the Tail code and the sound code (Fig. 3.5).

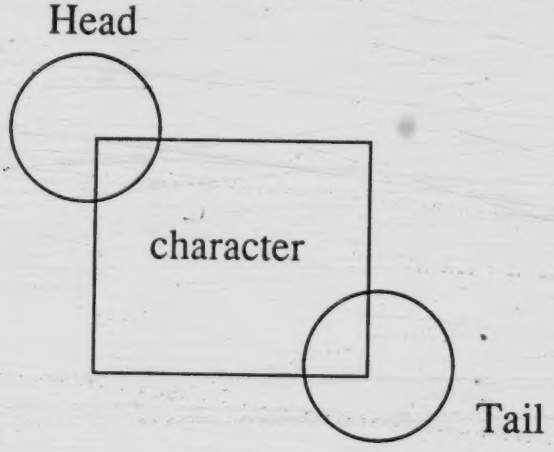


Fig. 3.5 The Head and Tail of a Chinese character

The Head code is the radical located at the upper-left corner of a character (note that a Chinese character is square in shape).

The Tail code is the radical located at the bottom-right corner of the character.

The sound code is the first letter of the Pinyin code for the character.

For example, for the character "琉" ("liu" in Pinyin)

Head code = 王

Tail code = 儿

sound code = L (the first letter of Pinyin code LIU)

In order to enter the Head/Tail code by a standard keyboard the keys have to be redefined to represent the radicals of Chinese characters.

For example,

the character "闷" (" wen " in Pinyin)

Head code = 门 = ";" key on the keyboard

Tail code = 口 = "Y" key on the keyboard

Sound code = "W" key on the keyboard

so, the user types " ; Y W" to get the character "闷 ".

Just as with the Pinyin code,the Head/Tail code is also a duplicate-code encoding method and a hint line is adopted for human-computer interaction.

For example, the Head/Tail code "2EH" will map to three characters "欢孩骇" and the characters will be shown on the hint line for a user to choose (by numerical keys). Figure 3.6 shows how the Head/Tail code works.

The Head/Tail encoding has three levels of indexing (Fig. 3.7). The first level list contains 36 bytes which represent 36 radicals of the code (i.e. 26 letters from A to Z, 5 numerical keys from 1 to 5, plus other 5 keys "space . , / ;"). The second level list is the indexing list for pointers to Head/Tail duplicate-code groups with each entry containing a pointer to the first duplicate-code of a group. The third level list is sound-code indexing list with one byte per entry. All Chinese characters have their own entry on the list. The entries are sorted according to Head/Tail code. The last list is the GB code list with 2 bytes per entry which stores the GB code of the character corresponding to an entry on the third level list.

The Head/Tail code works in the following way:

According to the first level list the correctness of a Head/Tail code entered by a user is checked. Then based on their position on the first level list the pointer to an entry of the second level list can be calculated. Next, from the entry of the second level list the pointer to an entry of the third level list can be found. Also, according to two adjacent entries on the second level list, the number of duplicate-code for a duplicate-code group on the third level list is obtained. The third level list is for checking the correctness of a sound code and for obtaining the number of duplicate-code of sound code. Finally, based on the entries of the third level list the corresponding GB code can be found from the GB code list and the hint line for human-computer interaction can be built based on these GB codes. The structure of Head/Tail code indexing lists is shown in Fig. 3.7.

Figure 3.7 illustrates an example of how to find a Chinese character (with the Head/Tail code as "AXB") through the indexing list.

Assume a user enters a Head/Tail code "AXB" (or AX_ if he does not know its pronunciation), then the first step is to check the first level list to find the position for the Head code A and the Tail code X, and then calculate the pointer to an entry on the second level list by the formula

(Head_position * 36 + Tail_position) * 3

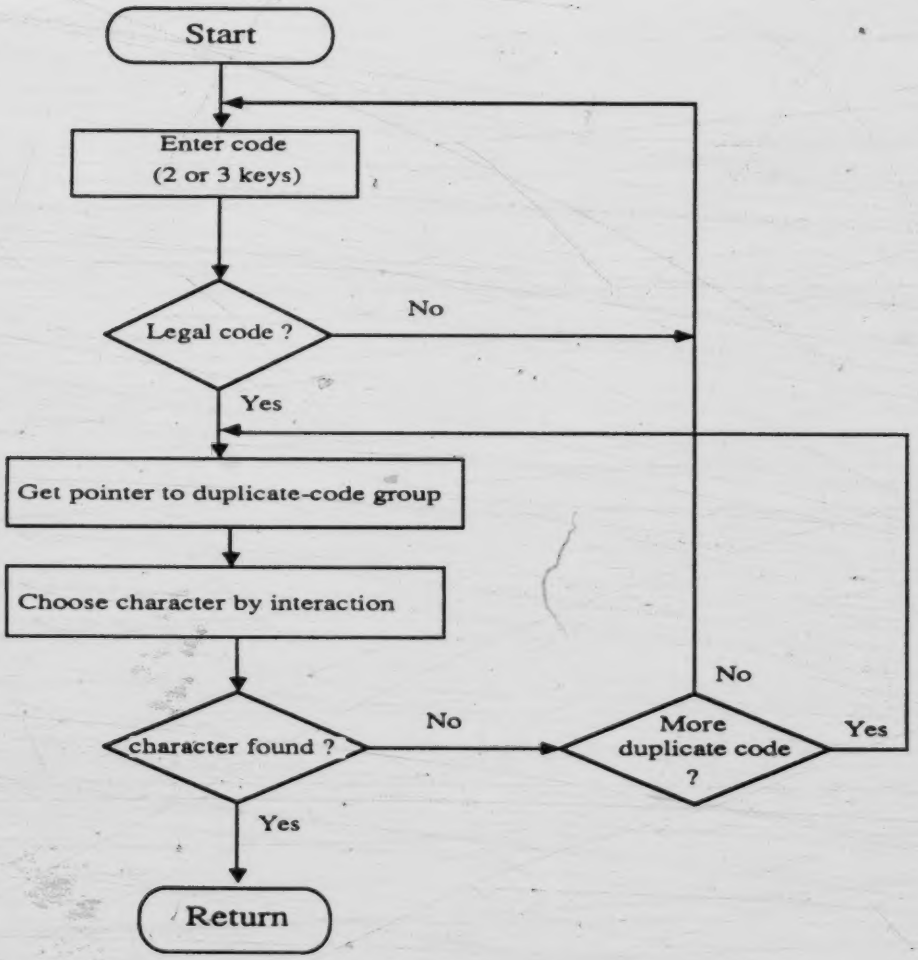


Fig. 3.6 Head/Tail Code Operation [GW84]

Next, find the content of the entry, N1, on the second level list, which points to an entry of the third level list (i.e. the first duplicate code of the duplicate-code group) and get the number of characters in the duplicate-code group by N2 - N1. Finally, find the GB code of the characters from the GB code list and show the characters represented by these GB codes on the hint line for the user to choose the character wanted.

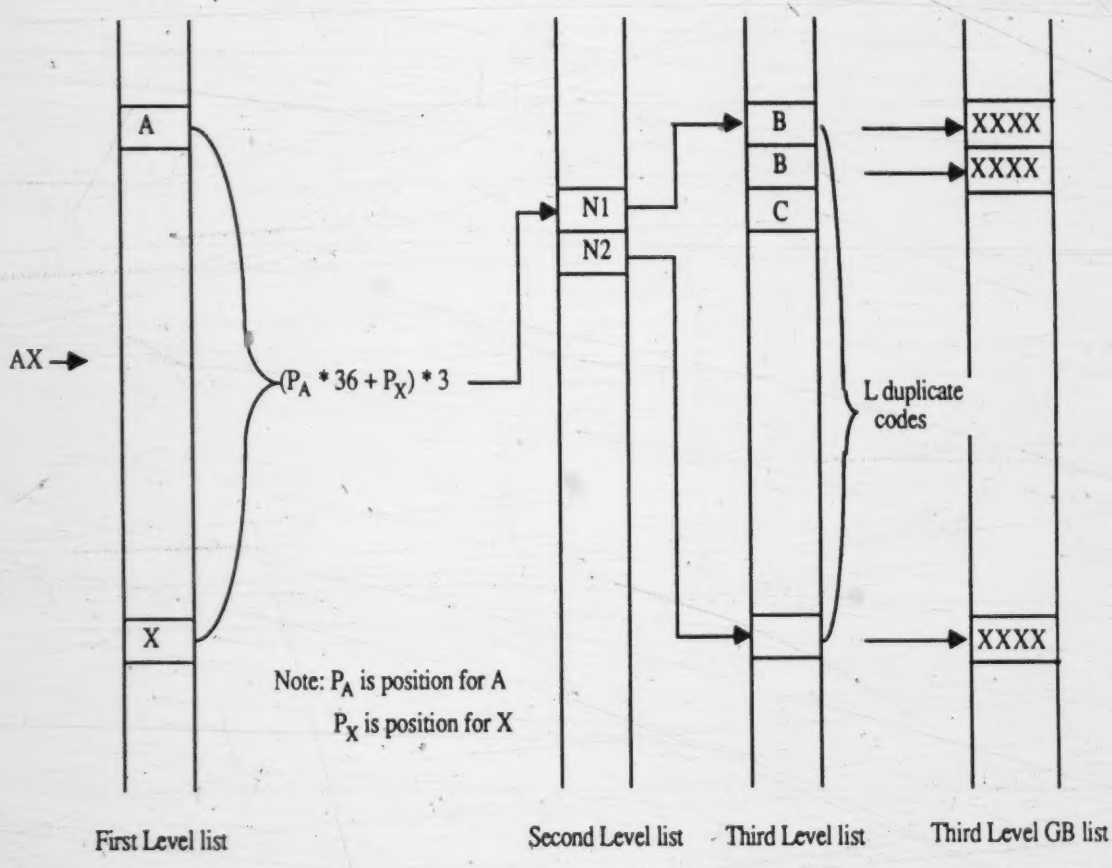


Fig. 3.7 The structure of Head/Tail code indexing lists [GW84]

Note: if the user does not know the pronunciation of the character and enters AX_, then all L=N2-N1 duplicate-code characters will be shown on the hint line. On the other hand, if AXB is entered (i.e. the sound of the character is known) only two characters are on the hint line (note that there are two "Bs" on the third level list which means two characters map to the same Head/Tail code AXB). Obviously, the sound code greatly reduces the size of the duplicate-code group.

3.4 Output Method

The computer generation of Chinese characters comes in several forms, namely, the softcopy display output on a CRT monitor screen, and hardcopy output on a printer. Just like western alphabets, Chinese characters have to be generated first by either fetching the prestored font or creating the font on-line before the characters are output at their destination.

In this section the features of Chinese output codes (i.e. the special requirement for the code and the methods to deal with this requirement) and Chinese font libraries are discussed.

Before proceeding, a point worth mentioning is that many think that input is the most important part of Chinese computing. Actually output is at least as important for the following reasons:

- "Seeing is believing", it is an output method that explicitely shows the difference between English and Chinese computing,
- Chinese output is as difficult as Chinese input,
- output is related to an important concept, treating text and graphics in an uniform way, which is so important to Chinese computing (and multi-lingual processing) that it is emphasized many times in* the thesis.

3.4.1 The Features of Chinese Font

The Chinese character set is very large and even 6763 characters of GB2312-80 is far less than enough. Although some characters are seldom used, they must be available in some way (either generated on-line or prestored in a font library) because in Chinese writing any character (font) cannot be represented by other symbols or information. In other words, when a user writes Chinese and encounters a seldom-used character, the character font must be found (e.g. by dictionary)and symbols cannot be used to represent the character (which is the way of Chinese writing). This is one reason for the huge character set.

The complicated shape is another feature of Chinese characters in that there is no ideal character-composition rule. Although each character is composed by five basic strokes, their positions, sizes, and directions vary from character to character. The same problem exists with character-composition by radicals in that there is no uniform rule to identify radicals. For example, the character "里" can be viewed as a radical or as a character composed of "田" and "土".

The advantage of a Chinese character is its square-shape, that is, all characters have the same size and no matter how many strokes a character consists of, they are all put into squares with the same size. Every character can be treated by a computer as an equal-sized image.

3.4.2 Font Representation

There are three ways to represent Chinese fonts in computer systems: dot matrix, skeleton vector, and contour vector (outline). Fig. 3.8 shows the character "上" ("Shang" meaning up) represented by these three ways.

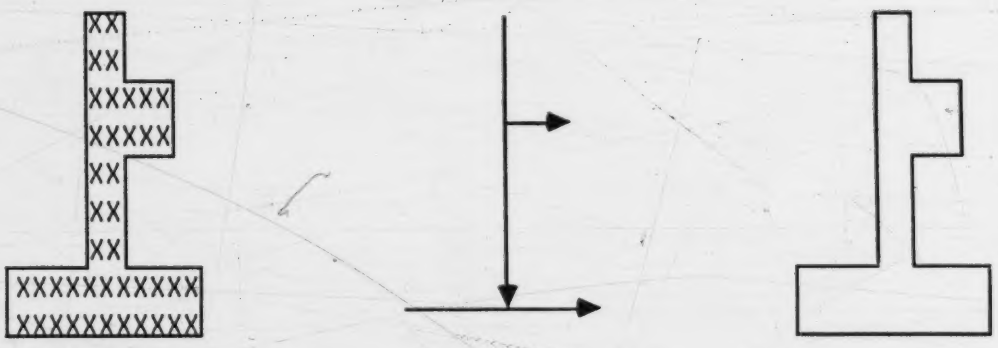


Fig. 3.8 Three ways to represent Chinese font

A dot matrix and contour can represent any Chinese character style. A skeleton often represents line-style character. For the memory space overhead, a dot matrix needs more space than the other two methods and the skeleton vector needs the least space among the three methods. When generating fonts, the dot matrix can generate them directly without any auxiliary technique, while a contour vector needs a "filling" technique, and a skeleton vector needs a "drawing" technique. So a dot matrix is the fastest one among the three.

According to different output requirements, the font library can be classified in three categories: low quality line characters, popular characters, and high quality character font libraries. The dot matrix and skeleton vector can be used for the first one, dot matrix is also used for the second one, and a contour vector and dot matrix are used for the third one, i.e. the high quality font library. So a dot matrix is good for all three categories.

A dot matrix is independent of the structure of Chinese characters, so it can easily represent non-Chinese characters and symbols such as graphics and table symbols. Whereas skeleton and contour vector methods depend highly on the structure of the Chinese characters. Therefore a dot

matrix is very useful for processing text and graphics in a uniform way (note: this is an important feature of Chinese computing which will be discussed further later). On the other hand, to generate high quality character with affordable space cost, contour vector is a very useful one (see chapter 6).

3.4.3 Digitizing Chinese Font and Storage Requirement

As discussed previously, prestored Chinese characters in dot-matrix form is the most popular way to output Chinese characters. For each font style and font size, each Chinese character will be digitized from its written form into a dot-matrix form usually as a square matrix (Fig. 3.9).

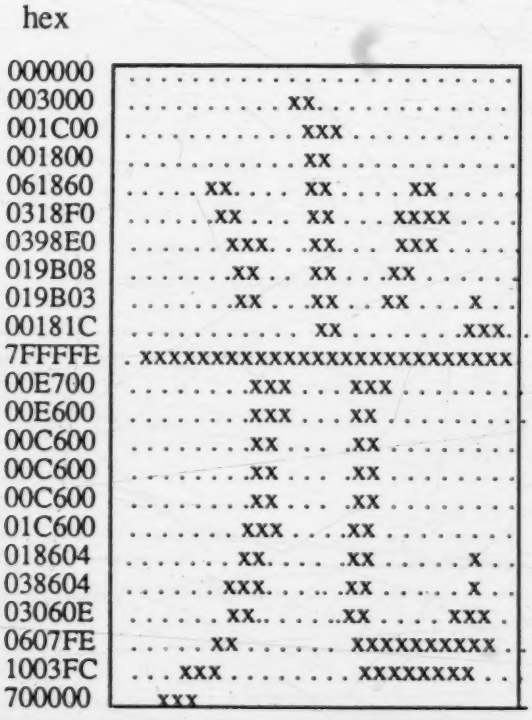


Fig. 3.9 The 24*24 dot-matrix for the Chinese character "光" (light) [ZHANG84]

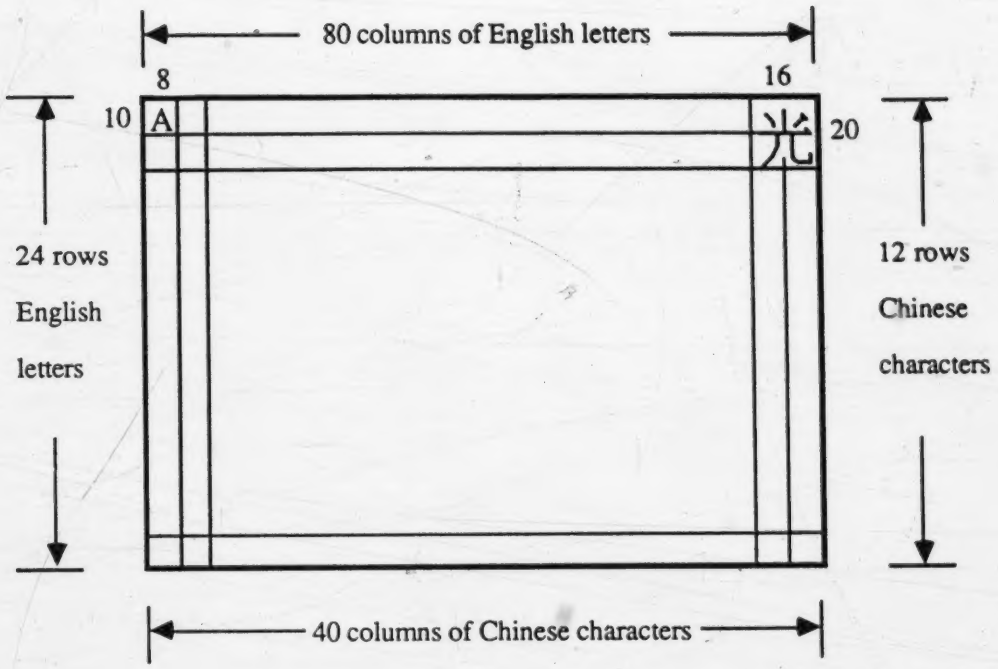


Fig. 3.10 To Display Chinese and English characters on an ordinary screen [ZHANG84]

A character with a 16 * 16 dot-matrix can be represented by 32 bytes (i.e. each row contains 16 bits or 2 bytes and there are 16 rows). Similarly, a character with a 24 * 24 dot-matrix needs 72 bytes. An English letter can be represented by a 8*8 dot-matrix which is only 1/4 of the space of a Chinese character (16 * 16 dot-matrix). Therefore, an ordinary screen with resolution of 640 * 240 pixels can display 24 rows of English text with 80 letters per row and 12 rows of Chinese characters with 40 characters per row (Fig. 3.10). To display high quality Chinese characters, a high resolution screen (e.g. with 1024 * 1024 pixels) is necessary.

The storage capacity of the Chinese font library depends on two factors: the number of Chinese characters and the size of the dot-matrix. According to frequency usage statistics of Chinese characters, the first level 3755 characters of the GB standard can cover more than 99.9 % of the characters used. A font library with 6000 to 8000 characters can satisfy most application of Chinese information processing systems. The storage capacities of the Chinese font library in common use are listed in Table II following.

Table II. The storage capacities corresponding to dot-matrix and number of characters [GUO85]

Dot-matrix	Number of bytes per character	Storage for 1024 characters (KB)	Storage for 4096 characters (KB)	Storage for 8192 characters (KB)
16 *16	32	32	128	256
24 * 24	72	72	288	576
32 * 32	128	128	512	1024

Note that the difficulty of output is due to not only the various sizes of the characters but also the variety of character styles. Because of historical reasons, there exist many styles of Chinese characters such Zhuan (篆), Li (隶), Kai (楷), Song (宋), Xing(行), Cao (草), WeiBei (魏碑), etc. Chinese calligraphy is not only an art but also a writing form which is widely used in publishing and daily life. This makes output of Chinese even more complicated, which will be further discussed in Chapter 5 and 6.

In order to generate high quality characters (or at least acceptable characters) with reasonable space overhead, many techniques are invoked which include font storage reduction by image compression/decompression, prestoring of basic components, Chinese meta fonts (generation by vector graphics), enlarging and smoothing from prestored smaller size characters, etc. [YU88]. However, due to the limitation of the thesis, they are not discussed.

3.5. Conclusion

In this chapter, information interchange code, input methods, internal codes, and output methods for Chinese are introduced, that provide the foundation for Chinese computing.

Chinese information interchange codes are usually established by government, e.g. GB 2312-80, which affects all other codes. Choosing a proper internal code is a central issue of Chinese computing systems in the sense that Chinese characters are represented as internal codes and processed by computers, and internal codes have great influence on Chinese/English compatibility. Two-byte internal codes are most popular in both China and the world. Input methods are the interface between users and computers. The main issue of input methods is how to map thousands of Chinese characters onto the standard keyboard. Two typical input codes are

introduced: Pinyin (input characters by sound) and Head/Tail code (input characters by shape). Three output methods are illustrated: dot matrix, skeleton, and contour vector. The space-cost for a dot matrix font library is analyzed.

Note that the following chapters are closely related to the issues discussed in this chapter. For example, both Chapters 4 (System Software) and 5 (Chinese Word Processing System) deal with internal codes and input/output methods. Chapter 6 (Chinese Electronic Publishing) mainly deals with output methods. Hence, this chapter is actually the foundation of the thesis.

Chapter 4
SYSTEM SOFTWARE

The topic of Chinese/English system software and the compatibility problem will be discussed in this chapter. This is one of the most important parts of the thesis. The Chinese/English compatibility problem is the central issue of Chinese computing. It is so important that whichever application (e.g. word processing, database, network, etc.) is intending to use Chinese, the first problem to be dealt with is the Chinese/English compatibility issue. Unfortunately, few papers discuss this problem and this chapter is intended to fill the gap. The discussion will focus on

- the importance of software compatibility.
- an analysis and comparison of two typical Chinese/English systems software for micro-computers (i.e. CC-DOS and 中文 Talk).
- the systematic approach to Chinese/English compatibility and the relation between graphics based computers and Chinese processing.

4.1. The Importance of Compatibility

For historical reasons, computers were invented in an English-speaking country, the United States, and most system software and applications software (including all high level languages) are written in English. Until now, English-speaking countries are still in the leading position of computer science and technology. Therefore, if the functionality of Chinese processing is added to an English computer system, then a great deal more versatility is available with the combined system in terms of the variety of software that can be used. Furthermore, and probably most important, future systems and software will probably see their greatest advances in English, which means that in order to keep up to date a compatible system is necessary. Another fact about the importance of Chinese/English compatibility is that Chinese people often use Chinese and English together. For example, when studying science, they use A, B, C, D, etc. instead of 甲, 乙, 丙, 丁 (a Chinese ordinal number system). A Chinese computer system which uses Chinese only is unacceptable by Chinese users. In addition, the computer industry is growing very rapidly in China.

Most computer systems developed in China are Chinese/English compatible systems and they are exported to Western countries such as European countries and the United States. In this sense, Chinese/English compatibility is necessary for people in the world to share the computer resources developed in China. For all the reasons mentioned above, the second approach has achieved great success among Chinese microcomputer system designers.

This chapter will focus on analysis and comparison of two important Chinese/English system software for microcomputers, CC-DOS for the IBM-PC and 中文Talk for the Macintosh. Because Chinese/English compatible microcomputers are widely used by Chinese people and CC-DOS and 中文Talk are successful systems software, conclusions on the language compatibility problem may be drawn by analyzing them. A brief survey on Chinese system software for minicomputers and mainframe is given at the end of the chapter. It is obvious that if the language compatibility problem can be solved in a systematic way, then computers will play a much greater role in all countries of the world.

4.2. CC-DOS, a Chinese/English Compatible System Software for the IBM-PC

Starting with this section, some available Chinese/English compatible systems will be introduced and evaluated. For each system, a brief introduction is given, then its compatibility is analyzed in the light of three aspects: the input-compatibility, the processing-compatibility and the output-compatibility. In other words, a Chinese/English system should allow users to input both Chinese and English, should be able to process Chinese/English mixed information, and to display the information on a screen or to print it out appropriately. Finally, conclusions are drawn for each of the systems.

4.2.1. A Brief Introduction to CC-DOS

The Chinese System CC-DOS (Chinese Character DOS) on the IBM-PC was developed in 1984 by the No. 6 Research Institute of the Ministry of Electronics Industry in China. Any IBM-PC and its compatible machines can run CC-DOS without hardware modification. The only thing a user needs is a floppy disc which stores CC-DOS and the Chinese font library. With the support of CC-DOS application software such as WordStar, Lotus 1-2-3, dBase II and III can use both Chinese and English characters with slight modification. These applications software are widely used in China now.

CC-DOS is based on MS-DOS (or PC-DOS), the system software for IBM-PC microcomputers. MS-DOS has the advantage of a simple and clear structure and based on this structure CC-DOS has good compatibility (but it is not as good as 中文Talk, which is discussed

later).

4.2.2.The Method Used to Extend MS-DOS to Accept Chinese

MS-DOS adopts a hierarchical structure (Fig. 4.1) which consists of three modules and one boot-program. The three modules are: Input/Output system (BIOS), file system (IBMDOS.COM) and command processing program (COMMAND.COM). BIOS is the lowest level of MS-DOS. It consists of a keyboard management module, a screen management module, a print management module, etc.

The simple hierarchical structure of MS-DOS brings great convenience to the development of CC-DOS in that each module in BIOS has a public interface and the higher level software always call the module through its public interface. This implies that a Chinese/English system can be developed by replacing BIOS by a Chinese emulation of BIOS. This is what the CC-DOS developers have done, and the structure of CC-DOS is shown in Fig. 4.2 .

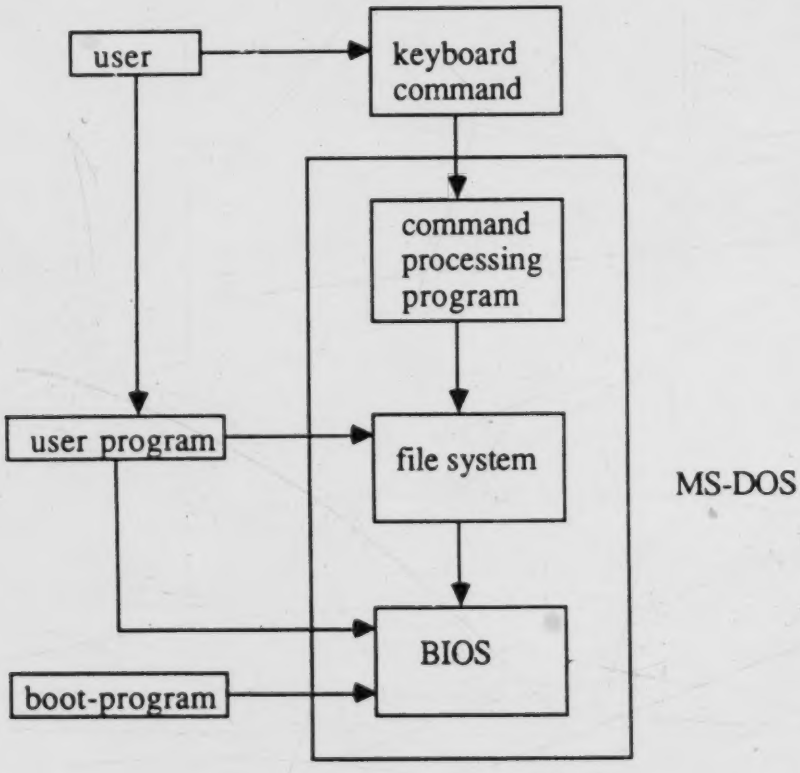


Fig. 4.1 The Hierarchical Structure of MS-DOS

On the other hand, it is shown by Fig. 4.1 that BIOS in MS-DOS can be accessed by application software (e.g. a user program) directly, which gives rise to a serious problem in developing a compatible system. There exist many kinds of applications and they may access BIOS in many ways. In order to achieve Chinese/English compatibility CC-DOS developers have to understand all aspects of BIOS thoroughly so that they may know every possible way in which an application software may access BIOS. Based on this information they may design a Chinese emulation of BIOS. Obviously, the task is too difficult and this may be the reason why CC-DOS does not obtain a good Chinese/English compatibility as is expected. This problem will be discussed further later.

The following three aspects of the compatibility of CC-DOS are described:

- how to achieve input compatibility
- how to achieve processing-compatibility
- how to achieve output compatibility

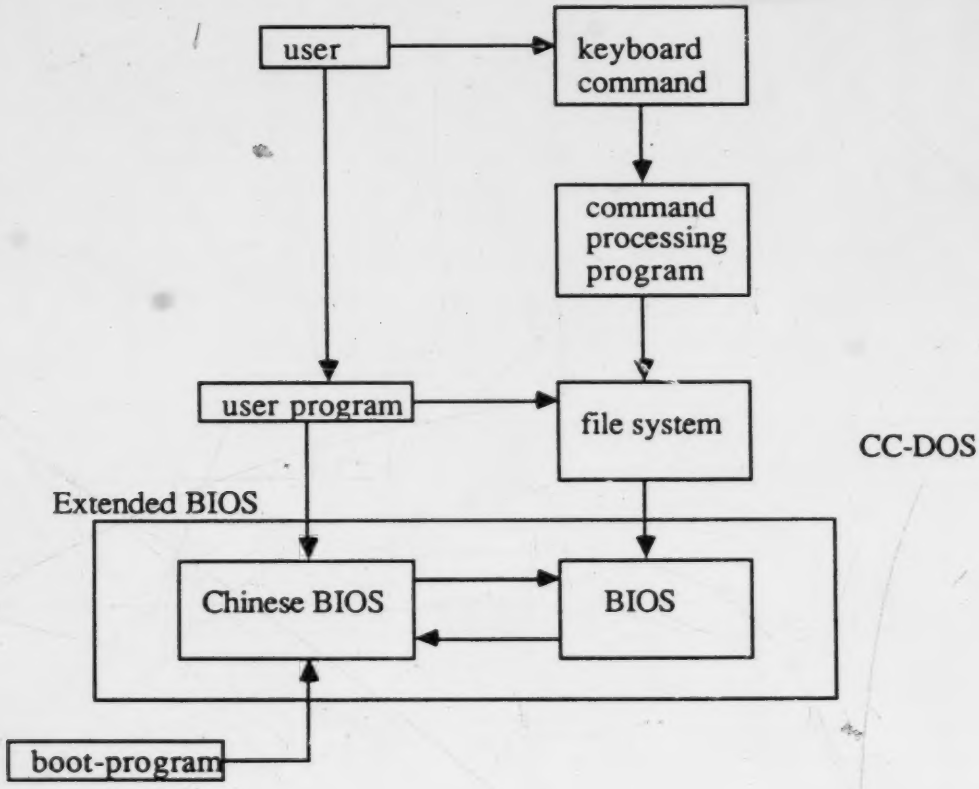


Fig. 4.2 The structure of CC-DOS

(1) Input Compatibility of CC-DOS

The input compatibility of CC-DOS is realized by the following steps:

- replace interrupt trap INT 16H by a new Keyboard management module so that when DOS asks for an input character it will call the new Keyboard management module instead of the original one. Fig. 4.3 shows the structure of the Keyboard management module of CC-DOS.
- distinguish English and Chinese characters by an Input code management module that stores a character from the keyboard in its buffer first, and then recognizes the character as a ASCII character, or a Chinese character and puts the character into the AX register .
- process the input code (input character) by the corresponding Input method module, that is, users inform the Input method management module which input method is being used (by F1 to F6 function keys on keyboard) so that the Input method management module can call the relevant Input process module (e.g. Pinyin input processing module) to process the input code.

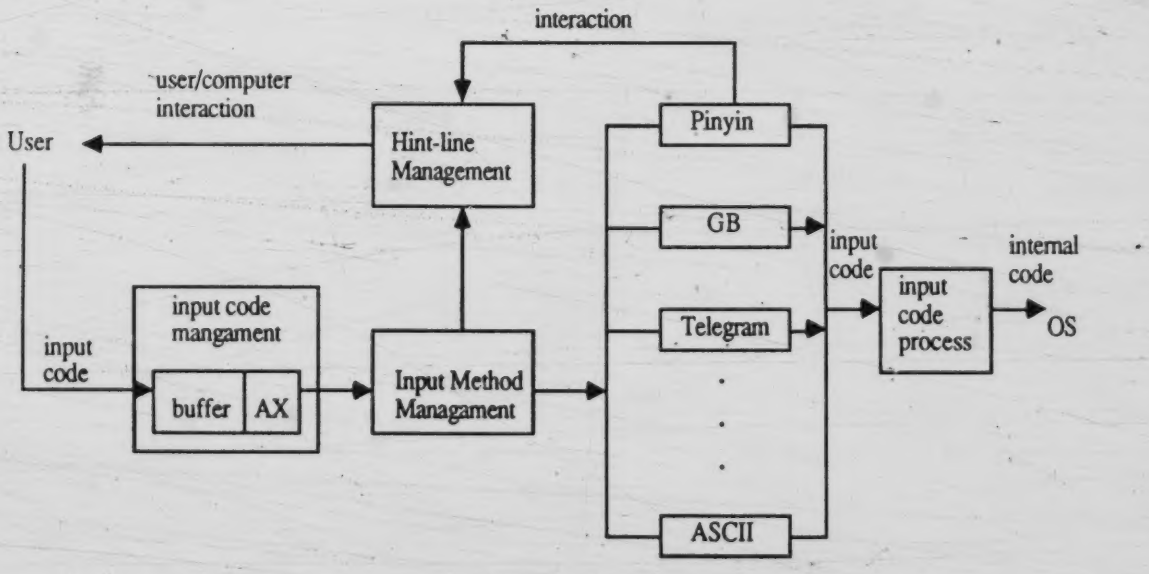


Fig. 4.3 The Structure of Keyboard Management Module of CC-DOS

- convert the input code (e.g. Pinyin code) to an internal code (i.e. ASCII code for English

characters or 2-byte code for Chinese characters) by the Input code process module.

- interact with the user by the Hint-line management module to solve the duplicate-code problem for some input methods such as Pinyin and Head/Tail code.

It is clear that the input-compatibility of CC-DOS is realized by the keyboard management module which separates Chinese input codes from ASCII codes and converts the input code to Chinese internal code.

(2) Processing Compatibility of CC-DOS

To process both English and Chinese characters properly inside a computer system, the key issue is to choose the proper internal Chinese code. In Chapter 3, various internal codes have been introduced and the principle based on which internal code is chosen was also given. In CC-DOS two internal Chinese codes were chosen. They are a two-byte code and a four-byte code (see 3.2). The two-byte code is the GB code with its 8th bit set to 1 (may be called meta-GB code). Since most software for the IBM-PC accepts the byte with its 8th bit set to 1 as a character, the two-byte internal code is widely used in Chinese computer systems such as IBM-PC, Great Wall 0520, and Macintosh. It is also convenient for editing, sorting, and storing. However, the two-byte code cannot be used as a communication code because in that case the 8th bit is assigned as a parity bit. In this case the four-byte code is used which contains three alphanumerical bytes (converted from the GB code) with a flag byte as the header to indicate a Chinese character. In short, an appropriate internal code will lead to good processing-compatibility. Since it does not conflict with the ASCII set, most available software will accept it. Also an appropriate internal code is easy to edit, sort and store [GW84].

(3) Output Compatibility of CC-DOS

The output of Chinese characters includes displaying the characters on a screen and printing them on a printer. In this section displaying Chinese characters will be discussed in depth because this part is the most sophisticated part of CC-DOS (in fact Chinese computer scientists ran into tremendous difficulties when designing the language compatible display module). Based on the discussion of output compatibility (together with the same discussion for the Macintosh in the following section) a very important conclusion will be drawn, that is, graphics-based computers (e.g. the Macintosh) provide a good foundation for the development of Chinese/English computer systems (as well as other language compatible systems), whereas text-based computers (e.g. the IBM-PC) lack such foundations to support language compatibility.

(3.1) Displaying English character

There are two display modes in MS-DOS: text mode and graphics mode. In text mode only

English text is displayed on a screen, while in graphics mode both text and graphics can be shown. Graphics mode is adopted in CC-DOS to display Chinese/English mixed text because of the pictorial nature of Chinese characters.

Fig. 4.4 shows the text mode of MS-DOS which illustrates that in text mode the CPU writes ASCII code into RRAM (regenerate-RAM) and then CRTC (CRT controller) recognizes the code and sends it to the character-generator to generate the character. Also the cursor is generated by CRTC. Text mode is not suitable for Chinese display because the character generator is too simple to generate a large numbers of Chinese characters.

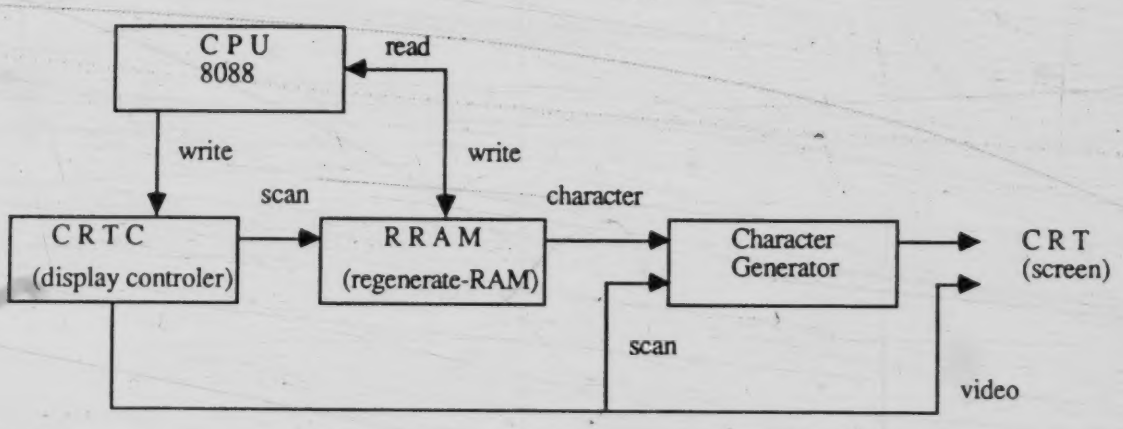


Fig. 4.4 The Text Display Mode of MS-DOS [GW84]

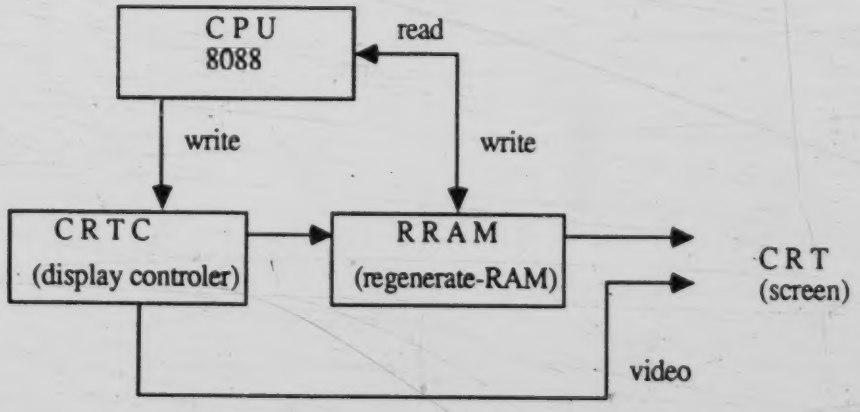


Fig. 4.5 The Graphics Display Mode of MS-DOS [GW84]

Fig. 4.5 shows the graphics mode of MS-DOS where the character generator is removed. In this case, if English text is to be displayed on screen, the text is treated as graphs and written into RRAM directly. Chinese text display can be achieved in the same way.

(3.2) The Difficulty of Displaying Chinese/English Mixed Text

Fig. 4.4 and 4.5 show that there is no character generator if the computer works in graphics mode. The character image is written into RRAM by the CPU directly. Then the character image (Chinese or English) will be displayed on screen. Besides, CRTC does not generate the cursor in graphics mode and the CPU has to write the cursor into RRAM. So, graphics mode (which is needed for displaying Chinese) is very different from text mode.

One major problem is that the number of lines for Chinese is 11 on the IBM PC screen while for English it is 25. This is because ASCII characters are displayed as 8*8 dot matrices while Chinese characters are displayed as a 16*16 dot matrix. Most application software of the IBM PC are based on the configuration of 25 lines of English text, so a mapping from 25 lines of English (ASCII) to 11 lines of Chinese is needed in order to achieve Chinese/English display compatibility.

To set the cursor properly on the screen is another major problem (actually it is one of the most difficult problems faced by Chinese computer scientists when developing CC-DOS [GW84, QIAN85]). Because the length of the internal Chinese code (2 bytes) is different from that of ASCII code (1 byte), the ways to process the cursor position have to be different. In English software, the cursor is moved and changed according to a basic unit (i.e. 1 byte), and obviously it cannot be used for Chinese without any changes. Furthermore, much of the software in the IBM PC can control the cursor position directly (e.g. BASIC language), which also complicates the display compatibility problem.

(3.3) Solution to Chinese/English Text Display: Virtual Screen and Real Screen

To display 11 lines of Chinese text on 25 lines of English-text-screen, a method called Virtual/Real screen is adopted (Fig. 4.6).

A virtual screen is a user-oriented buffer called VRAM (i.e. Virtual RAM) with 25 rows in it and behaves just like the ordinary English-text-screen. A real screen is the physical screen (which will display 11 rows of Chinese/English mixed text). The concept of a virtual screen makes it possible for a user to use the physical screen (11 rows now) as an ordinary screen with 25 rows of ASCII characters. This implies that English software supported by a 25-row-screen can still run using the support of CC-DOS CRTC.

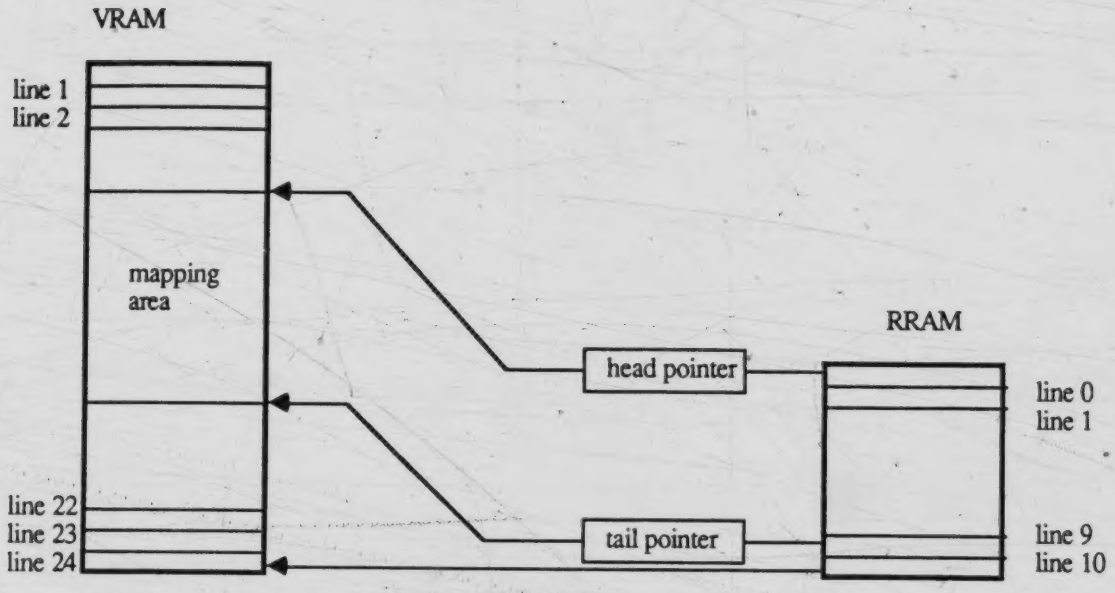


Fig. 4.6 The Relation between VRAM and RRAM [QIAN86]

The content of the real screen (i.e. the dot matrices of the characters) is stored in RRAM, the CRT Regenerate-RAM. While the content of the virtual screen (i.e. the 2 byte internal code for Chinese or 1 byte ASCII code) is stored in VRAM, i.e. the virtual regenerate-RAM, which contains 25*80 bytes to simulate the ordinary English-text-screen. The RRAM is related to VRAM by two pointers, a head pointer and a tail pointer. Since the virtual screen contains 25 rows while the real screen 11 rows, the text displayed on the screen is only part of the contents of the virtual screen. This part of VRAM (corresponding to RRAM) is called the RRAM-mapping-area and the two pointers point to its two ends (Fig. 4.6). The distance between the head and the tail pointer is always 9 rows which implies a 10 row mapping area (note that the last row of RRAM is the hint-line). By changing the values of the pointers, RRAM looks like moving within VRAM, which is called "the display-window".

A font buffer is added to CRTC to sort and process Chinese and English fonts and the content of the buffer is written to RRAM for displaying it on the screen. In this way, an English software applications can display mixed text on the screen.

(3.4) Solution to Cursor Generation and Location: Internal Cursor and External Cursor.

To solve the cursor-location problem a method called Internal/External cursor is adopted in

CC-DOS in which the external cursor is used to simulate and record the cursor-operation of software applications on VRAM (25 lines) and the internal cursor is used to generate and record the actual cursor-operation on RRAM (11 lines). Fig. 4.7 shows how to locate a cursor on the screen for Chines/English text. A brief explanation is given below.

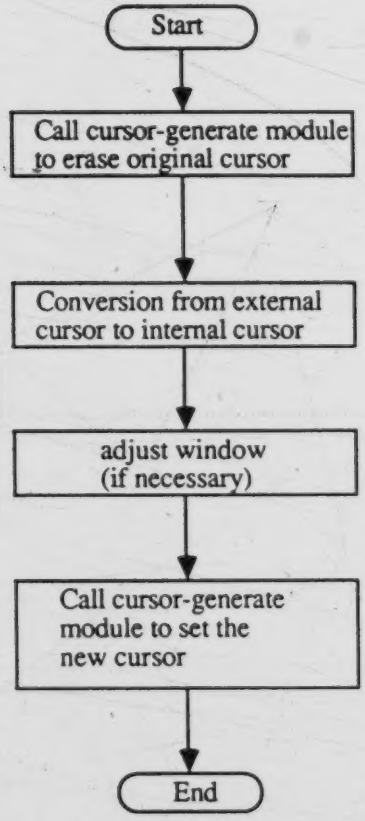

Start
Call cursor-generate module to erase original cursor
Conversion from external cursor to internal cursor
adjust window (if necessary)
Call cursor-generate module to set the new cursor
End

Fig. 4.7 The Procedure to Locate Cursor [GW84]

To solve the cursor-location problem it is necessary to

- Set the cursor on VRAM (external cursor) and the cursor on RRAM (internal cursor),
- Set the mapping between external and internal cursors,
- Create a cursor-process module to erase and draw a cursor.

In the Display-management module of CC-DOS, a cursor of an application software is represented as an external cursor and the head pointer and tail pointer are used to help locate the real

cursor on the screen (i.e. the internal cursor on RRAM). If a cursor to be located is not inside the display window, then the head/tail pointers are adjusted to locate it inside the window. The cursor-process module is called to erase the original cursor and draw a cursor at the new location according to a status-bit in graph-automatic-control. The status-bit can be set or unset from the keyboard or from TTY (TTY module is to display the current character and advance the cursor to next position automatically).

From the previous discussion, it is clear that a coherent problem with CC-DOS is that it has to work in graphics mode while the IBM-PC is not a graphics-based computer. Hence CC-DOS developers have to make great changes to MS-DOS and resolve all those difficult problems such as the mapping from the virtual screen to the real screen and locating the cursor properly.

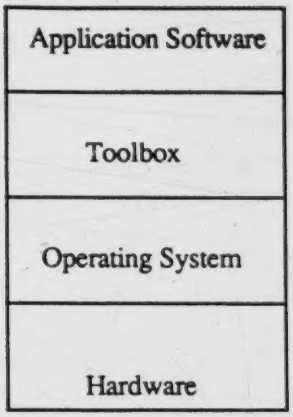
4.3. 中文Talk, the Chinese/English Compatible System Software for the Macintosh

4.3.1. A Brief Introduction to 中文Talk

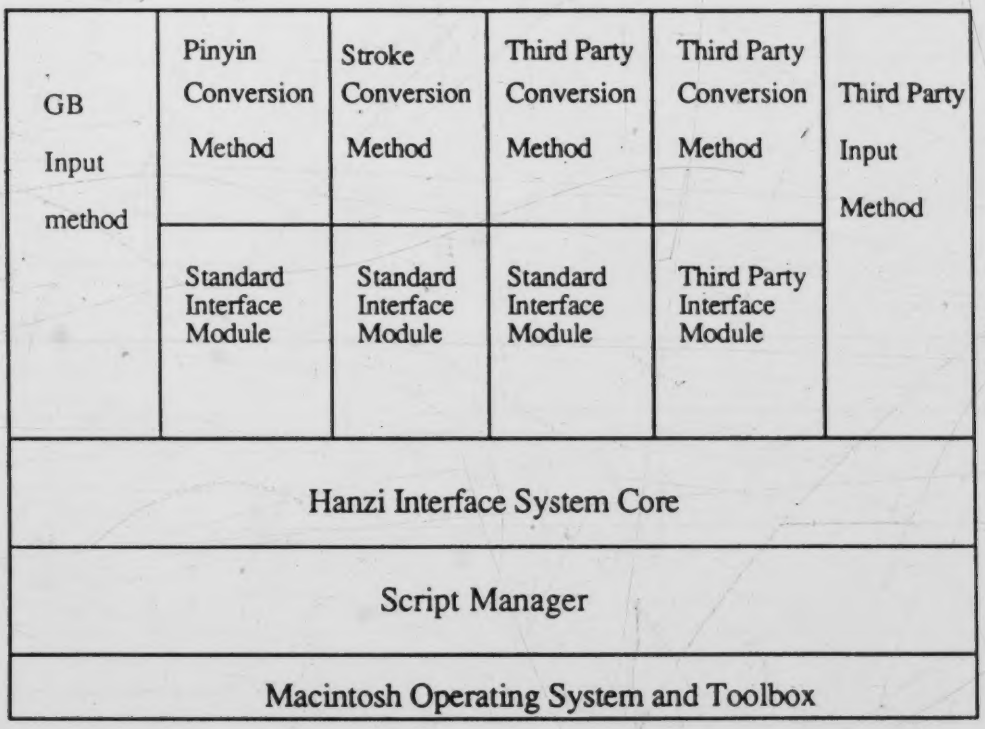
中文Talk, which is also called HIS (Hanzi Interface System), was released by Apple Computer Inc. for its Macintosh series microcomputer in June 1987. As mentioned in Chapter 1, a systematic approach is taken to develop the HIS, therefore, it obtains very good Chinese/English compatibility so that most Macintosh applications software can process Chinese.

Just as with CC-DOS, 中文 Talk stores 6763 Chinese characters (i.e. the two classes of characters in the Chinese Government Standard GB 2312-80). Users can use different methods to input Chinese characters (through the keyboard and mouse). Its input-methods-interface allows users not only to develop a new input method by themselves, but also to share the functions of dynamic retrieval and rearranging the order of Chinese characters. The software also allows users to add and modify character patterns in its font library.

The Macintosh system structure and 中文 Talk (HIS) architecture are shown in Fig. 4.8. Above the ToolBox, Apple developed a Script Manager, a HIS (Hanzi Interface System), and an Input Method Interface. The ScriptManager is for the languages other than English such as Arabic, Chinese, etc. The HIS includes a modified QuickDraw (for display), a modified Event Manager (for all input means such as keyboard, mouse, disk, etc.), a modified TextEdit module, and a Chinese font library (16*16 dot matrix for a character).



(a) The System Structure of Macintosh



(b) The 中文 Talk (HIS) Architecture

Fig. 4.8 The original Macintosh software system and 中文Talk architecture [HIS87]

4.3.2. The Method Used to Extend the Macintosh System to Accept Chinese

Fig. 4.8 shows that the Script Manager, the HIS, and input method modules are added to the top of the operating system and the Toolbox of the Macintosh in order to accept Chinese.

The Script Manager is a set of extensions to the standard Macintosh Toolbox and operating system that does two things:

- It makes it easy to translate an application into another writing system (e.g. Chinese).
- It provides standard, easy-to-use tools for the sophisticated manipulation of ordinary text.

The Script Manager is low-level software that enables Macintosh applications software to work with such different scripts like Chinese, Arabic, etc. It includes utilities and initialization code to create an environment in which scripts of all kinds can be handled [HIS87].

The Script Manager provides the functions needed to extend Macintosh's text manipulation capabilities beyond any implicit assumptions that would limit it to Roman scripts. The areas in which these limitations appear are:

- **Character set size.** Large character sets, such as Chinese, require at least two-byte codes for computer storage in place of the one-byte codes that are sufficient for Roman scripts.
- **Writing direction.** The Script Manager provides the capability to write from right to left, as required by Arabic and some Chinese publication in Taiwan and oversea Chinese communities, and to mix right-to-left and left-to-right directions within lines and blocks of text.
- **Context dependence**. Context means that characters may be modified by the values of preceding and following characters in the input stream (e.g. Arabic).

In short, the Script Manager provides a language compatible environment not only for Chinese and English, but also for all kinds of languages in the world (up to 64 scripts) [INSV88].

In order for application software to use a particular script, a **Script Interface System** to support that script must also be present. In the Chinese system it will be the Hanzi Interface System (HIS), i.e. 中文 Talk. A Script Interface System typically provides the following:

- fonts for the target language (e.g. Chinese)
- keyboard mapping tables
- special routines to perform character input, conversion, sorting, and text manipulation
- a desk accessory utility for system maintenance and control

The Script Manager calls a Script Interface System to perform specific procedure calls for a given script. How a typical call (in this case, Pixel2Char) is passed from an application through the Script Manager to a Script Interface System and back is shown in Fig. 4.9.

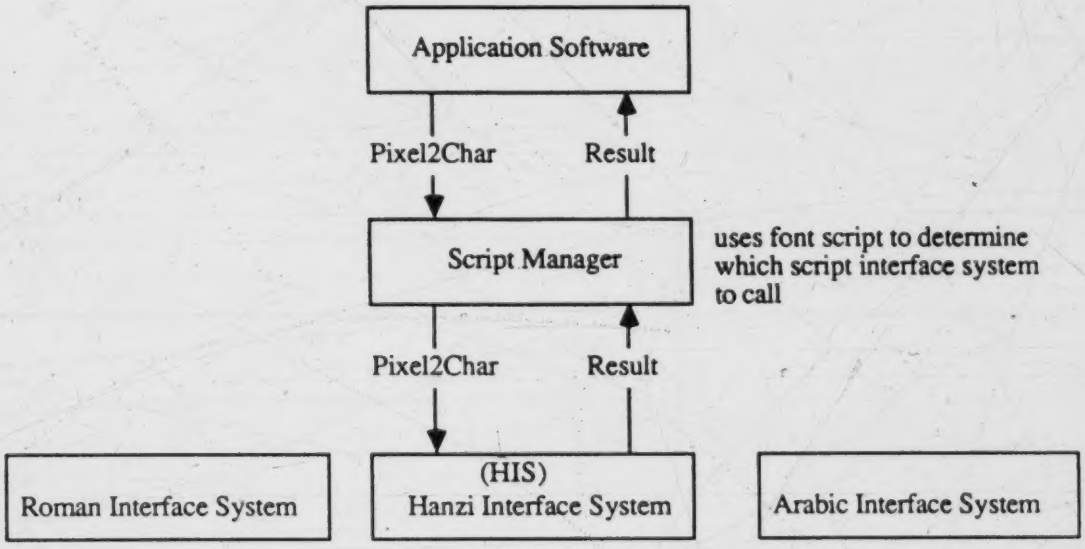


Fig. 4.9 The Script Manager guides a call from an application to the corresponding interface system [HIS87]

The HIS system is divided into two parts: the HIS Core, and the input methods. The HIS core modifies QuickDraw, Font Manager, and TextEdit to provide support for displaying and editing of Chinese characters. It also modifies Event Manager to support a "current" input method, and lets the user change the current input method through HIS utility calls (Fig. 4.8 (b)).

For input method developers, a standard interface module is also supplied. If the input method is a keyboard input method, the developer only needs to supply the conversion module, and links it to the standard interface module to create a new input method. Developers with a different user interface, such as a window for tablet-style entry, can write new interface modules compatible with the standard one, or create a completely new input method from scratch (Fig. 4.8 (b)). The following part of this section is the discussion of Chinese/English compatibility of HIS (i.e. 中文Talk), namely, how HIS supports applications software for input, processing, and

output of both Chinese and English information.

The lack of reference material makes it difficult to give complete details. For example, a main refrence is a technical report of HIS which is written for programmers not for system developers. The report simply provides programmers with a set of routines and tells how to call them if a programmer wants to write a Chinese/English application software. Therefore, it is difficult to understand the entire system and the relations among the routines. Particularly, it is not clear how Chinese/English characters flow through the system and how compatibility is achieved. In other words, the report does not indicate explicitely which modules take care of Chinese input, Chinese processing, and Chinese output respectively. It does not indicate through which interrupt traps the modules call each other. There is much less information for Chinese/English compatibility of 中文Talk than for CC-DOS. Even so, some features can still be explored which reveal the differences between CC-DOS and 中文Talk and some important conclusions can be drawn based on them.

According to the report [HIS87], the following modules are patched (based on original operating system and Toolbox routines) or newly developed to form the HIS Core:

- **Fonts module** - to display and measure Chinese characters, store Chinese characters in font resources, access and edit the characters.
- **TextEdit Support Module** - to handle 2-byte Chinese character during text editing, separate Chinese characters from English characters, find the location of Chinese characters on the screen, move insertion point properly.
- **Input Method Support Module** - to initialize input method, activate a current input method, respond to input events.
- **Script Manager Utilities Module** - all Script Manager Utility routines are included in HIS.
- **HIS Utilities Module** - to support code conversion, recognize Chinese characters in a string, create and maintain all tables for fonts module and input method module.

Fig. 4.10 shows some relations among the modules of HIS which are useful for analyzing the Chinese/English compatibility of HIS.

Note that the figure is not accurate in the sense that it does not include all details of the relations among HIS modules because of the complexity of the HIS structure and the shortage of references. The figure illustrates the general configuration of HIS. Besides, more emphasis is put on HIS Core because the Chinese/English compatibility problem is mainly related to this part. For example, in order to emphasize the HIS Core, TextEdit is considered as a module in HIS Core

although some routines in TextEdit belong to the Toolbox. Also, the HIS Ornament module is actually related to almost all other components of HIS, yet there are only two major links in Fig. 4.10 (i.e. the relations between HIS Ornament and Input Methods and between HIS Ornament and Fonts) in order to make the figure simple and clear.

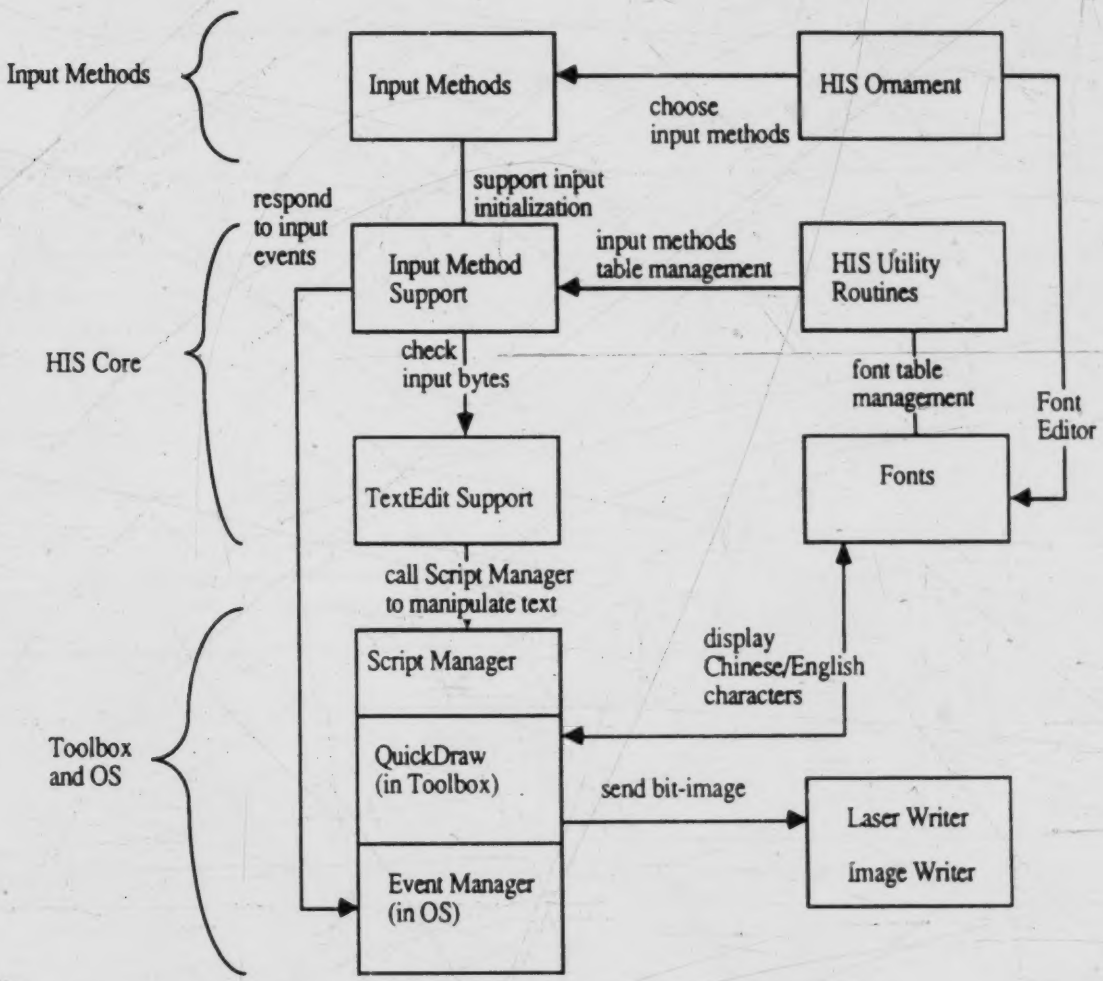


Fig. 4.10 Some relations among the modules of HIS

Note that most routines of the operating system and Toolbox are stored in ROM which implies that they cannot be modified by the HIS developers. So the author believes that the HIS developers extended the Macintosh system in the way similar to the CC-DOS way, i.e. modify the interrupt pointers to the area in RAM where HIS is located so that whenever the Chinese

information is encountered the system will pass control to HIS (in RAM) to process the information.

The explanation about figure 4.10 is given below together with the analysis of Chinese/English compatibility of HIS.

(1) Input Compatibility of HIS

The input compatibility of HIS is mainly related to the Script Manager Utility module, the TextEdit module, and the Input Method module. When a software application is running under 中文Talk (HIS), an icon symbolizing the script in use appears at the right side of the menu bar. The Chinese/English character recognition is performed in the following way:

When a user clicks the icon, i.e. to change the current script (e.g. English) to the other script (e.g. Chinese) this event informs the system and the **FontScript** field of the current **grafPort** is set corresponding to the script chosen (e.g. FontScript = 0 implies English and FontScript = 2 implies Chinese). (The grafPort is a data type in QuickDraw which represents the drawing environment). By checking the FontScript of **grafPort,** the system is able to recognize Chinese/English characters from the keyboard [INSV88]. If the FontScript indicates that Chinese is the current script, then the Hanzi Interface System (HIS) is called to process the Chinese character, otherwise the original Roman Interface System is called to process the English character. The input code conversion is achieved by the Standard Interface Module in the Input Methods Module. The standard interface module supports all of calls made by HIS on an input method and also defines a set of calls that conversion modules (which use the standard interface module) must support. (Fig. 4.8 (b)).

(2) Processing-Compatibility of HIS

As mentioned in Section 4.2, the key issue to process mixed Chinese/English text is to choose a proper internal Chinese code. In HIS a 2-byte Shift GB code(SGB) is used as the internal Chinese code which is different from the 2-byte internal code of CC-DOS. The internal code of CC-DOS has the simplest relation to GB code, i.e. it just sets the 8th bit of both bytes to 1, whereas the relation between SGB of 中文Talk and GB is not that simple.

The SGB is also defined based on GB code, however, the shift-difference between the SGB codes and GB codes is not the same for all Chinese characters, which complicates code conversion.

For example,

for the Chinese character "啊", GB = 3021

SGB = 889F

shift-difference D = SGB - GB = 587E;

while for the character "薄", GB = 3121

SGB = 8940

shift-difference D' = SGB - GB = 581F;

Note that D and D' are not equal which implies that the SGB code for a character cannot be generated from its GB code by simply shifting a constant number. After examing the SGB codes for all 6763 characters in GB character table, it is clear that the relation between GB and SGB is based on four shift-differences: D, D1, D2, and D3.

i.e. D = 587E, D1 = -5F, D2 = 1, D3 = -A2.

The formula to convert GB to SGB is

if i = even,

SGB = GB + D + (i/2)D1 + (i/2 - 1)(D2 + D3) for the first 63 characters in row i; (1)

SGB = GB + D + (i/2)(D1 + D2) + (i/2 - 1)D3 for the remaining characters in row i; (2)

if i = odd,

SGB = GB + D + (i/2 - 1)(D1 + D2 + D3) for all characters in row i; (3)

Note: there are 87 rows and 94 characters per row in the GB code character table (Chinese Government Standard).

Example 1:

the character "病" is on row 3 and its GB code is 3221, then

SGB = GB + D + (i/2 -1)(D1 + D2 + D3)

= 3221 + 587E + (2 - 1)(-5F + 1 - A2)

= 899F

Example 2:

the character "薄" is among the first 63 characters of row 2 and its GB code is 3121, then

SGB = GB + D + (i/2)D1 + (i/2 - 1)(D2 + D3)

= GB + D + D1

= 3121 + 587E - 5F

= 8940

Example 3:

the character "编" is not in the set of the first 63 characters of row 2 and its GB code is 3160 then

SGB = GB + D + (i/2)(D1 + D2) + (i/2 - 1)D3

= GB + D + D1 + D2

= 3160 + 587E - 5F + 1

= 8980

Obviously, the conversion between SGB and GB is not as simple as the one between GB and the internal code in CC-DOS. This factor has a negative effect on all code conversions in HIS such as the conversion between input code and SGB, the conversion between SGB and communication code.

In addition, there are some mistakes on page 13 of [HIS87] where a set of HIS Utility routines (which perform conversions among different codes) are given. GB2Index and SGB2Index are routines which convert GB (or SGB) code of a character to the index pointing to the corresponding font (dot matrix for the character):

Routine	From	To	Formula
GB2Index	GB	Index	Index = $((GB_{high} - 21) * 5E) + (GB_{low} - 21)$
SGB2Index	SGB	Index	Index = $((SGB_{high} - 81) * BD) + (SGB_{low} - 40)$

According to the formula the GB and SGB code for a particular character will be converted to different index code, i.e. a character has different fonts!

For example, the character "码" will obtain its font index as

GB = 3021 --------> index = 1410

SGB = 889F ------> index = 1398

Furthermore, since there are three conversion formulas between GB and SGB codes (see formulas (1), (2), and (3) above), there should be more conversion formulas from GB to Index and from SGB to index in the routines GB2Index and SGB2Index because the two simple formulas given above cannot guarantee that the GB and SGB code for a character will be converted to the same index of the font for that character.

In summary, for some unknown reason 中文Talk adopts a strange internal code which

complicates code-conversion. In addition, since the SGB code sets its two 8th bits to 1 as the internal code of CC-DOS does, the author believes that 中文Talk also uses the 8th bit as the Chinese character indicator.

(3) Output compatibility of HIS

In displaying Chinese/English characters on a screen 中文Talk of the Macintosh shows a great advantage over CC-DOS due to the fact that the Macintosh is a graphics based machine while the IBM-PC is a text based one.

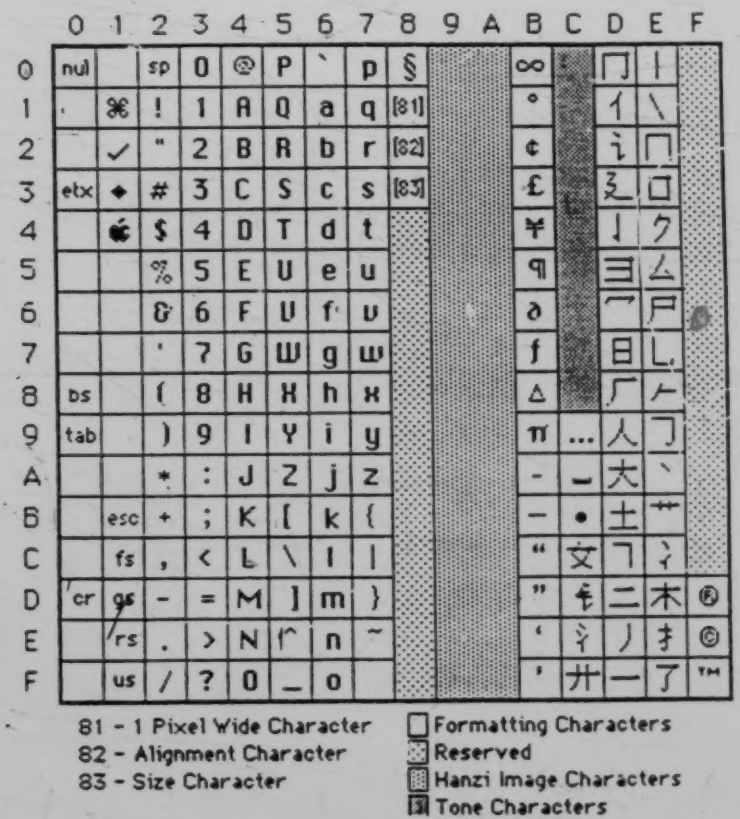

0 1 2 3 4 5 6 7 8 9 A B C D E F
0 nul sp 0 P p §
1 ⌘ ! 1 A Q a q [81]
2 ✓ " 2 B R b r [82]
3 etx ◆ # 3 C S c s [83]
4 $ 4 D T d t
5 % 5 E U e u
6 & 6 F V f v
7 ' 7 G W g w
8 bs (8 H X h x
9 tab) 9 I Y i y
A * : J Z j z
B esc + ; K [k {
C fs , < L \ l |
D cr gs - = M] m }
E rs . > N ^ n ~
F us / ? O _ o
81 - 1 Pixel Wide Character
82 - Alignment Character
83 - Size Character
Formatting Characters
Reserved
Hanzi Image Characters
Tone Characters

Fig. 4.11 The Chinese Font of the HIS (中文Talk).

The HIS (中文Talk) uses a Macintosh font for displaying Chinese characters. The fonts (Fig. 4.11) have 32 characters reserved, from 90 to AF, for drawing the Chinese characters. These positions are initially empty, and must be filled with the appropriate images before

QuickDraw is allowed to use the font. They also contain the stroke and tone characters that are used by the Chinese input methods.

When a string is drawn, it is first separated into blocks of either ASCII or Chinese characters (the routine FindBlock in HIS Utility Routine module will take care of this). Blocks of Roman characters are always drawn directly with the font (Fig. 4.11). Blocks of Chinese characters are divided into segments of 32 or fewer characters. For each segment, the image for each character in the segment is copied into the strike of the font. A string consisting of the bytes from 90 to AF is then printed using the font. This mechanism allows the full power of QuickDraw, including styles and arbitrary sizes, to be applied to Chinese characters.(Fig. 4.12)

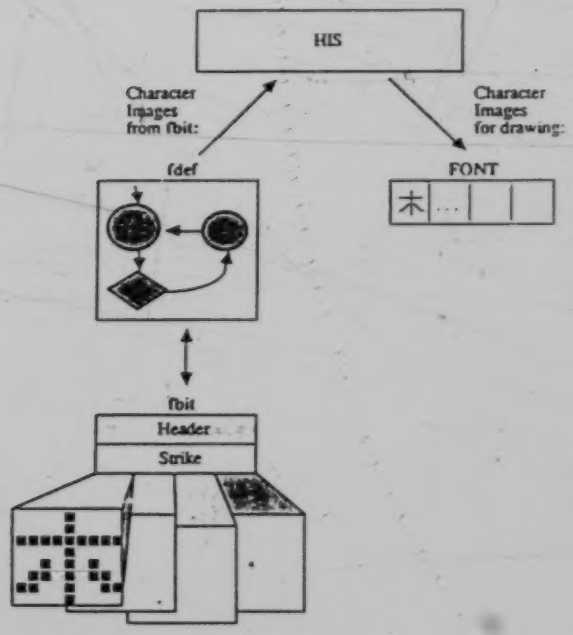


Fig. 4.12 Collecting and copying characters into the strike of the Chinese font.

A brief explanation of Fig. 4.12 is given below:

fbit - a resource type where the Chinese characters are stored and an **fbit** consists of a fixed header followed by unspecified data (the character image). The header contains the information about the font such as the resource ID number of the **fdef** used to

manipulate this font, the height and width of the character images in bits, the associated Macintosh font family, the font style and bit-depth of the character images (used for higher-quality fonts), the SGB codes of the first and last characters in this **fbit** resource.

fdef - a resource which contains several routines to retrieve a character image from **fbit**, or store a new character image in it. Since all access to the character images is funneled through the **fdef** resource, HIS does not need to know the format of the character data, and different Chinese fonts can be stored in different formats as long as they are accompanied by a **fdef** that knows the format.

When HIS is installed, it loads font resources from the HIS file itself and any font files it can find. It then builds tables to keep track of the installed Chinese fonts. These tables determine which font will be used to draw Chinese characters, depending on the font family (e.g. Beijing), size (e.g. 16*16), and style (e.g. bold) values of the current GrafPort.

It is clear that the difficulties encountered by CC-DOS developers such as the mapping from 25 lines ASCII characters to 10 lines of Chinese characters on screen simply do not exist in the Macintosh because the Macintosh is a graphics-based machine and those problems have been solved by its operating system and Toolbox routines (mainly QuickDraw, TextEdit, and Font Manager) already. A Chinese character is just put into a font (Fig. 4.11) then QuickDraw will do all the rest such as drawing the character with different styles and sizes for 中文Talk.

4.4. The Comparison of CC-DOS and 中文Talk

A comparison between CC-DOS and 中文Talk is necessary due to the fact that many are concentrating on implementing Chinese computing at the popular PC level, the majority of Chinese computer systems available are at this level, and CC-DOS and 中文Talk are typical ones among them.

4.4.1. The Similarities of CC-DOS and 中文Talk

Two approaches have been taken by the developers of the Chinese computing systems, namely, by software only and by hardware and software. Low cost and readly availability are the advantages for the software approach.

(1) Both CC-DOS and 中文Talk use the software approach which implies that the only thing a user needs to use Chinese on a microcomputer is a disk (or two disks) which contains the Chinese system and the Chinese font library. The user can simply boot the Chinese system and the font

library into main memory and then use Chinese without requiring any extra hardware.

(2) Both systems aim at high degree of Chinese/English compatibility (although 中文Talk achieves much better compatibility than CC-DOS does).

(3) Both systems adopt a double-byte internal code based on GB code (although CC-DOS does provide four-byte codes for users as the secondary choice).

(4) Both systems work in graphics mode in order to display and print Chinese characters. More precisely, CC-DOS requires an IBM-PC computer with a graphics card in it and the Macintosh is a graphics based computer by its own (i.e. it always works in graphics mode).

4.4.2. The Differences between CC-DOS and 中文Talk

(1) The structures of CC-DOS and 中文Talk are different. 中文Talk is built on the Script Manager which is an international interface sitting on the top of the operating system and Toolbox of the Macintosh. The Script Manager supports not only Chinese text manipulation but also other languages such as Japanese and Arabic. Whereas CC-DOS is developed from MS-DOS (the operating system of the IBM-PC) directly and there is no international interface sitting between CC-DOS and MS-DOS.

(2) CC-DOS has a very clear hierarchical structure which is related to the structure of MS-DOS, whereas the structure of 中文Talk is complicated which is related to the complicated structure of the operating system and Toolbox of the Macintosh. This may be one of the reasons that CC-DOS is the most popular Chinese system in China and many software applications are written for it, while the 中文Talk is almost unknown to Chinese users.

(3) In displaying Chinese characters on a screen, 中文Talk has great advantage over CC-DOS in that displaying Chinese characters is an easy and natural matter because of the graphics-oriented nature of the Macintosh. While displaying Chinese characters is a completely new task for the IBM-PC due to its text-based nature.

(4) 中文Talk achieves much better Chinese/English compatibility than CC-DOS does. Note that there may be two levels of compatibility, namely, low and high level compatibility. At the low level, an English application software can run on a Chinese system, but it can not use Chinese. At the high level, an English application software accepts Chinese under the support of the Chinese system. Most applications software in the IBM-PC obtain only low level compatibility under CC-DOS, while many applications software in the Macintosh obtain high level

compatibility under 中文Talk.

The reason for 中文Talk's good compatibility is that:

(1) 中文Talk is developed by Apple itself, while CC-DOS is developed by third party developers. It is difficult for the developers to understand MS-DOS thoroughly without enough technical references.

(2) Many English applications software fail to display Chinese characters on a screen because of the text-oriented nature of the IBM-PC, while 中文Talk has few problems with the task.

In summary, CC-DOS is a Chinese system software with good structure so that developing an application software on it is easy, but CC-DOS does not achieve a good Chinese/English compatibility as expected. In contrast, 中文Talk obtains good compatibility, but its structure is complicated so that developing an application software on it is difficult.

Finally, people may ask "Is there any way to develop a Chinese system with both clear structure and good Chinese/English compatibility?" This will be discussed in the following sections and the section Character Channel in Chapter 8.

4.5. Systematic Approach to Develop a Chinese System

In Chapter 1, two approaches to the development of Chinese system were briefly introduced. Two points were emphasized for the systematic approach which are (1) a solid foundation (i.e. hardware and operating system) which is suitable for Chinese Processing; (2) a standard which brings three groups (i.e. computer manufacturers, third party software developers, and users) together to achieve a highly compatible system. Based on the detailed analysis of CCDOS and 中文Talk, the approach may be described clearly now.

Some important observations on developing a compatible operating system are the following which lead to the concept of the Systematic approach:

(1) So far there has been no systematic way to expand an existing English software so that it can use Chinese. The computer scientists at the Chinese Academy of Science just input Chinese characters to an application running under a Chinese operating system (e.g. CCDOS) and trace the flow of the characters. Usually the characters seldom go through every part of the software smoothly and if they are blocked somewhere, computer scientists try to find the location and the reason, and then modify the software to force it accept Chinese characters. Sometimes, a number of good features of the original English software have to be sacrificed in order to make the software work with Chinese.

(2) A compatible system is usually an extension to the original English system and therefore much more complicated. Surprisingly, people who develop new systems seldom provide users with easily understandable documents. For example, CCDOS was developed by the No.6 Research Institute of Electronic Technique of the Ministry of Electronics in 1984 in Beijing, however, it was Dr. Qian Peide from SuZhou University who gave the detailed analysis and flow-chart of CCDOS in his book [QIAN86] . This required considerable effort to understand MS-DOS and CCDOS of the IBM-PC. The book is one of the most important references for Chinese Computing.

For the Macintosh 中文Talk, the document [HIS87] provides mainly a set of routines for programmers to use and it is difficult to understand how Chinese and English characters can be used in the new system without conflict. As a result, a flow chart of 中文Talk is given in the previous section.

(3) The shortage of information exchange between Chinese users and system developers leads to the deficiency in developing a good Chinese/English compatible system.

Obviously people in China are the main users of Chinese computing systems and Chinese computer scientists (e.g. the CC-DOS developers) clearly know the requirements of the Chinese/English compatibility a computer system should have, but lack the technical details of computers like the IBM-PC and the Macintosh (which makes it difficult to develop a compatible system). On the other hand, a computer company like Apple owns all documents discribing the technical details of its machines so that it is most capable of developing a compatible system. On the other hand, programmer analysists in a computer company may not understand Chinese and the status of computing in China very well (although the company may hire Chinese specialists to work for it). Maybe that is why 中文Talk is not as successful as it could be. For example, the graphics-feature of the Macintosh implies that 中文Talk can provide many excellent Chinese character styles (e.g. Li style, Kai style, etc.) which will make 中文Talk a very successful Chinese system in both China and the world. Unfortunately, it does not make full use of the feature. Also, the conversion between SGB code and GB code is not very simple which will have a negative effect on all aspects of Chinese processing.

Some important conclusions can be drawn from the observations:

(1) Cooperation between three groups (i.e. computer manufactures, third party software

developers, and Chinese users) is necessary. For example, Apple Computer needs help from Chinese software developers in China, because the latter know more on Chinese and CHinese computing than the former. And the latter is able to help Apple to design a better internal code and to make full use of the graphics features of the Macintosh. Another example is that if CCDOS developer cooperated with IBM and MS-DOS developer, i.e. they had enough technical details of the IBM-PC and MS-DOS, then the CCDOS developer would not take the Patching approach and CCDOS would be a better compatible system.

Chinese users are also a necessary member of the cooperation and should be well informed with information of the compatibility of a Chinese system (such as Character Channel of the system which will introduced in Chapter 8), because many users know computers, i.e. they may develop software for their own purposes. The software should be Chinese/English compatible. Additionally, if some of their applications have compatible problem under the Chinese system, they need a way to figure out the problem.

In short, in order to avoid the Patching approach and develop an integrated Chinese computing system, cooperation between the three groups is necessary.

(2) A standard is necessary which gives information on compatibility. This is the central part of the cooperation between the three groups. The standard should give information on character-flow in a computer system. Especially, the interfaces where characters pass through should be indicated explicitly. Information on code conversion is also required. A new concept, Character Channel (see Chapter 8), is based on this. The standard should be open to public just as ASCII code so that the three groups can follow it to achieve the best compatibility. For example, a third party software developer follows the standard to write a Chinese/English compatible application software (e.g. a software written according to the ScriptManager of the Macintosh is guaranteed to be Chinese/English compatible). A user follows the standard to solve his unique compatible problem, e.g. if he wants to set a network, he needs the code conversion information, and if he has compatible problem with an application, he can find out the problem according the standard.

(3) A solid foundation (i.e. hardware and operating system) is necessary for the development of an integrated Chinese system. Because of the nature of the Chinese language, graphics-based computers such as the Sun and the Macintosh have obvious advantages over texe-based computers. If CCDOS developer chose a graphics-based computer, they would not run into the great trouble with displaying Chinese characters on a screen. Next section will analyze this in detail.

4.6. Graphics-Based Computers are Suitable for Chinese Processing

Many feel that graphics-based computers is better than text-based computers in Chinese Computing, but they seldom inquire into the reason behind this. Based on the analysis of 中文Talk and CCDOS, the relation between graphics-based computers and Chinese processing can be revealed. In this section, major features of graphics-based computers are introduced first, then the reasons why the features are suitable for Chinese processing are given.

The major features of graphics-based computers are the following:

(1) A high resolution, raster graphics screen which offers flexibility for the display of arbitrary objects, including variable width characters, foreign alphabets, mathematical symbols, lines, curves, shaded regions, and photographic images [SUN86]. For example, the Sun Workstation uses a high-resolution (1152*900 or 1600*1280 pixels) 19-inch display monitor and the Macintosh uses 512*342 9-inch display.

(2) Powerful graphics processors and operations. For example, the Sun's Graphics Processor (PG) uses a two-stage pipeline and a floating point chip set, that performs transformations, clipping, scaling, and rendering operations on vectors and polygons. The PG also incorporates a Painting Processor to manipulate a color frame buffer and fill polygons. The PG improves the performance of graphics operations by a factor of 2 to 250 depending on the operation. Similarly, the Macintosh QuickDraw (also called Screen Manager) is a collection of over 100 drawing commands. It includes routines to draw lines, rectangles, ovals, polygons, and irregular shapes, outlining or filling them with a variety of pen sizes and patterns. QuickDraw is stored in ROM to obtain high speed [MORG85].

(3) Iconic system. Namely, small pictures (icons) are widely used to help users operate the computers. For example, both the Sun and the Macintosh use icons to represent file name, software, disk, and commands.

(4) Powerful window system which is combined with mouse to result in a more natural and dynamic communication between user and computer. For example, SunView (Sun Visual/Integrated Environment for Windows) allows multiple overlapped windows to be concurrently active on the Sun display. SunView includes Textedit (Mouse Oriented Editor), Iconedit (Icon Development) and Fontedit (Font Development). Similarly, the Macintosh window system is a major part of Toolbox (a system software supporting user interface) and the system includes FontManager and TextEdit to support various character operation.

(5) Drawing and painting software. Drawing software is to generate vector object such as lines, polygons, and curves. Painting software is for the raster graphics generation such as an image with different gray levels. For example, both the Sun and the Macintosh provide such application

(e.g. **fig** of the Sun and **MacDraw** and **MacPaint** of the Macintosh) and programming environment.

In short, the combination of high level hardware (for graphics) and high level user interface is an outstanding feature of graphics-based computers.

The features of graphics-based computers suit the requirements of Chinese computing for the following reasons:

(1) The high resolution, raster graphics screen can display Chinese characters with various styles and sizes. For example, there is no difficulty for the Macintosh to display Chinese characters, while the task is almost impossible for a text-based terminal of the IBM-PC.

(2) Powerful graphics processor and operations are foundations for generating and displaying high quality Chinese characters efficiently, which is a key issue of Chinese computing. Because of the pictorial nature of Chinese characters, almost all Chinese system works in graphics mode, which implies a much slower processing speed than in text mode. A slow processing speed makes any Chinese application software unacceptable by users. Hence, a graphics processor or accelerator is necessary for Chinese computing.

(3) Iconic and window system are foundations of a friendly user interface which is extremely important for Chinese processing because:

• Chinese input is a bottleneck of Chinese computing and hundreds of input methods are developed to help users input Chinese easily and efficiently. A friendly user interface provides a foundation for input methods. For example, to generate a hint-line of the Pinyin method, CCDOS developer has to consider the layout-arrangement of the screen, while for a graphics based computer, it is just a matter of opening a window. Also, CCDOS uses number keys to indicate a homonym, while the Macintosh 中文Talk has more ways to do it, i.e. by arrow key or by clicking mouse.

• To improve efficiency of a Chinese system, associative font library and phrase library are widely used. A typical example of associative library is that users enter one character, then computer displays a number of phrases and idioms with the character as starting character for users to choose from [NI85]. This technique is widely used in Chinese computing to improve Chinese input and editing. A graphics based computer can open multiple windows for users to choose the desired character and phrase.

• Character online generation is necessary when a desired character is not available in a font library. A multiple window facility can greatly ease the task. For example, one window is opened for drawing outline of a character with the other window displaying the filled character for users

to compare with.

• It is more difficult for ordinary Chinese people to use computers than for Western people, because the former do not have the tradition of using typewriter. In this case, a friendly user interface is even more important for Chinese people to get used to computers. For example, they do not have to remember many commands, to use keyboard heavily, or to worry about the operation modes (note: "modeless" is an primary design-objective for the Sun and the Macintosh designers).

(4) WYSIWYG (what you see is what you get) is an advantage of the graphics based computers and drawing/painting applications make it even more powerful. This benefits Chinese computing in many ways. For example, drawing/painting applications often have better Chinese/English compatibility than a word processor. If a word processor cannot work with Chinese properly, Chinese characters can be stored in a painting application software and then embedded in a file of the word processor.

(5) Graphics-based computers process text and graphics in integrated way. They provide font management facility that can be directly used for Chinese processing. For example, the Macintosh allows users to design a desired font. So, all the 中文Talk developer needs to do is to specify a new font (called Chinese font) and locate desired characters in it, then the Chinese font can share the full power of QuickDraw, e.g. the generation of characters with different artistic styles and sizes.

In short, the bottleneck of Chinese computing is Chinese input and output. Graphics based computers provide an environment powerful enough to solve the bottleneck problem. In other words, many facilities required by Chinese computing are available in Graphics-based computer systems already, while with text-based computers a Chinese system developer (e.g. CCDOS developer) has to design the facilities from scratch.

4.7. Other Systems

4.7.1. The Hardware/Software Approach in Microcomputers

Nowadays, more Chinese computing systems have adopted the dedicated hardware approach as technology advances. Hardware called Chinese Card (or 汉卡 in Chinese) are designed not only to modify the display for higher resolution but also to use the external RAM to store font libraries, dictionaries and input methods. In addition, it enhanced the computing power of Chinese computer so that it may compete with the original English system. The trade off is

performance for cost. Typical examples are Cang Ji Chinese Card (仓颉) for the Apple II and the Associated Chinese Card for the Great Wall 0520 (developed by Ni Guangnan in Beijing) [NI86].

4.7.2. Chinese Systems for Workstations

The implementation is quite popular in Japan in supporting the Japanese language (note: 2,000 Chinese characters are included in Japanese). However, it is not yet popular in Chinese language handling. Probably it is because of the higher cost of the workstation and also people are concentrating on solving the problems at the popular PC level. For example, the Xerox's Star workstation has multilingual support including Chinese. The team at the Institute of Systems Science of the National University has created a Chinese language environment on the TI Explorer system. Also, some work on creating a Chinese environment on IBM RT/PC workstation is also done at the Institute of Systems Science. The AT & T, the developer of UNIX system, is cooperating with the Institute of Computer Software in Beijing to expand UNIX into an international system. According to the People's Daily [PD89], an international UNIX system has been developed and the Institute is developing a Chinese subsystem based on the international system. This is another good example of developing Chinese system in systematic way.

4.7.3. The Chinese System Software for Mini/Main Frame Computers

For mini and main frame computer systems, it is not easy to modify the operating system. The Chinese computing is implemented with the use of the Chinese terminals, in most cases. By using certain coding techniques, the system may take Chinese data without modification. At this level, designs are come under three categories, software, hardware, and PC Emulation. At the software level, hanzification of the Operating System and the applications are required. At the hardware level, the terminals are modified to display Chinese character (with a Chinese Card), and the printers are able to print Chinese characters. At the PC Emulation level, it uses existing Chinese PCs with a terminal emulation program. For example, IBM 3270 PC Emulator for IBM 4300 series and IBM 9370.

4.8. Conclusion

In this chapter Chinese/English system software and the compatibility problem are discussed, which is one of the most important parts of the thesis. Two typical software systems, CC-DOS and 中文Talk, are analyzed in terms of Chinese/English compatibility (i.e. input, processing and output compatibility). The results indicate that a systematic approach is necessary for the

development of Chinese computing and graphics-based computers such as the Macintosh provides a good basis for achieving Chinese/English compatibility. A new concept, the Character Channel, is introduced in chapter 8, which tries to solve the compatibility problem in a systematic way.

Chapter 5
CHINESE WORD PROCESSING SYSTEMS (CWP)

CWP is the most important software application in Chinese computing due to the following facts: (1) CWP reflects two most outstanding features of Chinese computing: input and output of Chinese/English mixed text; (2) CWP challenges the central issue of Chinese computing, the Chinese/English compatibility, that is, if the compatibility problem can be solved for CWP, it can be easily solved for other software applications such as spreadsheet, graphics, etc. Therefore, CWP deserves one of the most important parts of the thesis.

Due to economic reasons, the big majority of the Chinese computer users today are using word processors that are easily accessible with low cost. Namely, the word processors initially written for English text editing were modified by Chinese system developers for editing Chinese text (some Chinese systems can use English word processors to edit Chinese text without patching the word processor's code). The popularly used CWP under this category are Chinese WordStar, PE, and PE II. As these word processors were originally used for English text editing, some features do not work correctly when it is applied to Chinese text [YU88].

There are, however, word processors that are specially written for Chinese text editing. More popularly used ones are: WANG Advanced Ideographic Word Processor, KAIHIN CWP (called BrushWriter), TIANMA CWP, etc. These CWPs will be studied because they show the full range of CWP features.

This chapter will first give brief introduction to three typical CWPs, i.e. KAIHIN, TIANMA, and 中文TALK, then detailed evaluation criteria for CWP and their justification are given, next, weights are assigned to the criteria, and finally the results of experimental tests are compared with the results of evaluations of KAIHIN, TIANMA, and 中文 Talk. Note that the evaluation by three methods, i.e. Wei's criteria, Lo's criteria and the criteria presented here, are outlined in Appendix A.

5.1. A Brief Introduction to Three CWPs

5.1.1. KAIHIN CWP

The KAIHIN CWP called BrushWriter was developed in 1986 by Kaihin Research Inc. in Canada. The system includes a Macintosh Plus, a printer controller and a 24*24 dot matrix printer. It can also form a multi-user typing system (note: the newly released version does not require the controller).

Four input methods are provided for Chinese character input. They are Pinyin, GB code, Row/Column method, and the Radical/Character-based method developed by KAIHIN Inc. A standard keyboard is used. The Pinyin method is the standard one, so are the GB code and Row/Column methods. For the radical method, 36 most commonly used radicals are arranged on the keyboard and they can form other complex radicals. A nice feature of the Radical method is intelligent error correction, that is, the system accepts out-of-sequence radical strings and can tolerate input errors to some extent [CHIN87].

The system makes use of the window features of the Macintosh which include menu and mouse operation, dialogue windows, online help, etc. Macintosh editing functions, such as cut, copy, paste and replace characters, are also included. Users are allowed to design a new character and put it into a Chinese font library. The Chinese font library contains 7,000 characters and can be increased to 14,000 characters.

The system needs an extra printer controller to output the characters with different styles and sizes. Unfortunately, those character styles cannot be shown on the screen and special symbols are inserted to indicate the appropriate character styles for the controller (and printer) to generate them.

5.1.2. 中文 TALK

中文TALK is not a Chinese Word Processor but a Chinese version of the Macintosh's operating system. In theory, if an operating system is changed to accept Chinese, all software application including word processors can use Chinese. In practice, however, some extra work is needed to help Chinese codes work with software applications smoothly. In June 1987, Apple announced 中文Talk in Beijing and that three software applications, MacPaint, MacDraw, and MacProject can use Chinese under 中文Talk. 中文Talk also supports MacWrite but many operations of MacWrite do not work correctly. Although 中文Talk is not a CWP, it shows the input/output features of CWP, and more importantly, a powerful CWP can be developed based on 中文Talk because it makes full use of the feature of the Macintosh. Therefore, it is also included in this chapter.

The main advantage of 中文Talk over KAIHIN is WYSIWYG (what you see is what you get), namely, the characters with different styles and sizes can be shown on screen. They can be printed out on an ImageWriter and a LaserWriter directly without extra hardware. 中文Talk provides four input methods, namely: Pinyin, BiXing (stroke/radical), GB code, and Rome, and it

allows users to connect their own input methods to the system. The Pinyin method is better than the standard one in that the homonym characters are sorted according to their usage-frequency, so the "hit rate" is high. 中文Talk stores 6763 characters (2 levels of Chinese characters in Chinese Government-Standard with 16*16 dot matrix for each character) and users are allowed to add or modify character patterns. 中文Talk's input-methods-interface has wide functions so that users can not only develop new input method by themselves, but also share the functions of dynamic retrieval and rearranging the order of Chinese characters.

5.1.3. TIANMA CWP

TianMa is developed by International Geosystems Corporation in Vancouver, Canada, in 1985. It works on the IBM PC or compatible computers. It consists of a TianMa system disk and a TianMa ROM card. Pinyin is the only input method in TIANMA. The great advantage of TianMa's Pinyin over other Pinyin methods is that user can type Pinyin text continuously. The build-in intelligence analyzes the context of a Pinyin sentence and translates it into Chinese automatically. Therefore TianMa CWP can accept text four or five times faster than other Pinyin methods because the other Pinyin methods have to have a decision from a display of up to 200 homonyms for every character.

Beside automatic conversion, TianMa has other features, such as Pinyin spelling help, self learning, phrase dictionary, and the Pinyin Spelling Tutorial. Users can update the character/phrase dictionary such as changing the order or add new phrases to it. If TianMa chooses a wrong character or phrase from a homonym list, the user can replace it quickly. If a user corrects the same wrongly selected homonym three times, TianMa will "remember" the correct one when the next time it comes across the homonym (i.e. self-learning). TianMa provides commonly used editing/formatting operations such as setting margins and tabs, word-wrap, centering, setting headers and footers, page numbering, etc..

For outputting Chinese, TianMa provides Simulated Song style only with two sizes (small character and large characters) and the user can also choose simplified or classical Chinese characters. All the above can be shown on screen or printed out on a printer. The printer can be any of the following parallel printers (or compatibles): IBM Graphics Printer, GTC Blaser, Epson LQ series.

Note: TIANMA 2+ (with more character styles) will be released sometime in 1989 and this information is not included here.

5.2. Criteria for Evaluation

The criteria for evaluation of a CWP consists of 4 parts (Tables III to VI): Input Method Criteria, Pinyin Input Criteria, Output Criteria, and Editing/formatting Criteria. There are 4 levels for each criterion, i.e. A, B, C, and D which stand for Excellent, Good, Neutral, and Poor respectively.

5.2.1. Interpretation of Input Method Criteria (Table III)

Unlike many input-method (or keyboard) evaluation approaches which concentrate on irrelevant single criterion such as duplicate-code rate or the number of strokes per character, the approach presented here attempts to evaluate an input method in its entirety. Namely, it examines whether users can concentrate on their writing (not the code conversion), and whether the input method provides good facilities to assist with input, etc. The argument for this approach is that a CWP is an integrated tool instead of a set of separate elements. Sometimes, a CWP is optimal for a single criterion (like duplicate-code rate), but the CWP is not always user-friendly.

Criterion 1: According with user's habits

Conversion is necessary for any keyboard input of Chinese. For the phonetic method the conversion is:

sound -> Latin alphabet -> typing.

For the component method (e.g. stroke or radical decomposition methods) the conversion is :

character -> stroke/radical -> typing.

This conversion makes Chinese input very different from English input. To input English, one merely inputs sequential characters or shapes, that is,

shape -> typing,

and it is not necessary to go through the phonetic (or shape decomposition) stage. A good Chinese input method should reduce the burden of the conversion. In other words, it should allow the user to concentrate on what is to be said or written instead of coding rules, key location, etc. [LO85].

Criterion 2: Coding rules

To input Chinese characters the first problem a user has to deal with is to remember coding rules of an input method. The simpler a rule is, the shorter the learning time will be. The learning time is one of the most important factors for judging an input method [WEI85].

Table III Input Method Criteria

No.	Criteria	A	B	C	D
1	According with user's habits, so that user concentrates on his thinking instead of code conversion.	allow user to type while thinking; conversion doesn't affect thinking; seeing characters knowing the code; very easily and smoothly typing, etc.	conversion affect thinking slightly, user can still concentrate on his thinking; easily and smoothly typing.	conversion affects thinking; typing is not very easily and smoothly.	concentrating on code conversion and typing instead of thinking.
2	Easy-remembered coding rules and short learning time (LT).	No need to remember special rules; LT = 0	the rules can be remembered by ordinary people in short time; LT = 1 week (approximately).	the rules can be remembered by ordinary people in long time. LT = 1 month (approximately).	the rule can be remembered by professionals only.
3	Duplicate-code rate: average number of strokes per character:	< 0.5 % 3.7	< 1 % 4	< 2 % 5	< 4% > 6
4	Input-method-interface.	user can easily install his own input method; provide more than 5 methods and can switch from one to another in any cases.	hard to install user's own method although it is possible; provide more than 5 methods and can switching is allowed in most cases.	cannot install user's own method; about 5 methods are provided and switching may not be easy.	can't install user's own method; only one or a few methods are provided.
5	Input speed and touch-typing.	< 200 Hanzi/min, (Hanzi means Chinese Character); touch-typing.	< 100 Hanzi/min touch-typing	< 30 Hanzi/min semi-touch-typing; need to look at keyboard in some case.	< 15 Hanzi/min no touch-typing
6	Size of keyboad and extra hardware needed.	standard keyboard; no extra hardware needed.	standard keyboard; need extra hardware (e.g. PC-graphics card).	middle-size keyboad; with or without extra hardware	large keyboard with or without extra hardware.
7	Facility and higher degree intelligence.	on-line help; self-learning-tutorial; associative ability (e.g. input of phrase, idiom, commonly used sentence structure, etc.) error tolerant.	part of facilities and intelligent functions in category A. e.g. on-line help self-learning-function,and error tolerant to some extent.	limited facilities and intelligent function e.g. on-line-help and self-learning only.	nothing to help user's input and the user has to know and remember all details.

Criterion 3: Duplicate-code rate

Duplicate code means a single input code may represent more than one character. Obviously, the duplicate-code-rate should be as low as possible. The rate chosen in Table III is based on the 6763 characters of the Chinese Government Standard Table (GB-2312-80).

Criterion 4: Input-method-interface

Usually an input method interface should be flexible and powerful so that it can support more than one input method. However, a user may use only one or two methods and all others just waste memory space and complicate usage (actually some methods like the telegraph code may never be used). The number 5 is chosen as a threshold due to the following consideration: a good interface should provide at least 1 or 2 phonetic methods (e.g. standard Pinyin and Cantonese Pinyin), 2 or 3 component methods (e.g. Wang's 5 stroke method and Zhi's "Seeing character knowing code method", etc.), plus the GB code. Switch-ability means to allow a user to switch from one method to another in any operation mode of in case one method cannot satisfy the user (e.g. the user is using Pinyin, but forgets the pronunciation of a character, then it is necessary to use a component method to input the character). A user's favorite input method may not be available in a CWP. In this case, the CWP should allow the installation of the method.

Criterion 5: Touch-typing

Touch-typing is the process of typing without looking at the keyboard. Usually touch-typing increases input speed. It is also good for "typing while thinking" because looking at the keyboard from time to time will interrupt a user's thinking. Therefore speed and touch-typing are put together.

Criterion 6: Keyboard and extra hardware

This criterion is about hardware considerations. Both specially designed keyboards and extra hardware will increase the cost of a CWP.

Criterion 7: Facility

Various facilities and some kind of intelligence are so important that a user can hardly use a CWP without them. For example, without on-line help, a user has to check manuals frequently. If the CWP is not error tolerant (to some extent), input may be stopped regularly due to typing mistakes, and if a CWP provides a new input method without a self-learning-tutorial attached to it, new boring rules may be difficult to remember.

5.2.2. Pinyin Input Method

The Pinyin input method will be studied carefully in this thesis and because of its advantage over component methods. The author believes that Pinyin will be the first National standard input method in China. Therefore the Evaluation criteria for the Pinyin method alone (i.e. Table.5.2) is given. The following is an overview about the Pinyin method.

1. The Advantages of Pinyin Method:

(1) Most Chinese people in mainland know Pinyin due to the fact that since 1958, Pinyin has been taught in elementary schools all over the country.

(2) Mandarin is the official dialect in both mainland and Taiwan, so even if some people do not know Pinyin, they can pick it up easily (Note: "Mandarin" is not the proper name of the official language. The proper name is "Pu Tong Hua", i.e. 普通话 in Chinese. However, "Mandarin" is the best way to express it in English so that most people will understand).

(3) Pinyin is taught in Chinese courses all over the world, that is, all foreign students who learn Chinese also know Pinyin.

The above 3 reasons lead to the following results:

- Pinyin input method is widely used in China although it is slow;
- Many oversea Chinese language experts strongly recommend Pinyin as a National standard input method. Many versions of Pinyin input method and CWP based on Pinyin are developed in China and abroad [TAINMA85, LIU87, JIANG87].

(4) Pinyin is standard, namely, it does not change with time. Whereas almost all component methods are changing (e.g. the well-known 5-stroke method by Wang Yongmin has changed three times) and among hundreds of component methods none of them have been accepted as the best for 15 years competition (in 1974, Chinese government started its first Chinese information processing project). The basic reason for this is that there is no standard rule to decompose Chinese characters by shape. In addition, any component method is harder to learn than Pinyin because Chinese characters are very complicated ideographic symbols and it is difficult to convince ordinary people to learn a complicated non-standard input method.

(5) Pinyin is especially suitable for CWP users. When a user is using CWP the natural way is to type while thinking and the text appearing in his mind is mainly its sound instead of the character's shape. If the user is copying something from available text he has to think about the character's shape. However this is not the common case of using a CWP.

(6) Pinyin is a possible way for indexing just like English. Most Chinese know when looking up Chinese dictionary it is much easier to use Pinyin than radicals. The reason is that Chinese characters are not sortable by their nature and there is no standard way to decompose or construct 60,000 Chinese characters. By the Pinyin method characters are divided into 414 groups (by Pinyin syllables) then each group is further divided into 4 sub-group (by 4 tones) and finally the characters in a sub-group are sorted by the number of strokes or some other ways.

2. The Disadvantages of Pinyin Method.

The Pinyin method does have problems, and it will not become a National standard input method unless these problem are solved.

(1) Homonym.

Since thousands of characters are mapped onto 414 syllables, there exist the average of 9 homonyms in the first level of the Chinese government standard (3755 characters). This leads to a slow input speed because for each character the user has to indicate the character from a list of homonyms.

(2) Dialect.

Pinyin is based on the Mandarin dialect which implies if a user does not speak Mandarin (like Cantonese people) he can not use the Pinyin method. Considering most dialects in southern China are very different from Mandarin, Pinyin does have serious problems with this.

(3) Unknown characters.

Pinyin is used under the condition that user knows the pronunciation of the characters. If a user wants to input an unknown character, he has to look up dictionary to find its sound and then input it by Pinyin. Looking up a dictionary will slow down the input speed. Chapter 8 contains the detail analysis on these problem and the possible solutions which

will not be repeated here. The following is the interpretation of Table IV, the evaluation criteria of the Pinyin input method.

5.2.3. Interpretation of Pinyin Input Criteria (Table IV)

The idea behind the criteria is to emphasize smooth typing (because Pinyin typing is usually interrupted by homonym-indication), to emphasize phrase-input (because phrase-input can reduce the size of homonym-list significantly), and to emphasize a higher degree of intelligence (because it is essential for solving 3 main problems of the Pinyin Input method described above).

Criterion 1: Continuous-typing

Continuous typing is the main goal that a good Pinyin method should achieve. The common reason for users not to accept the standard Pinyin input method is its lacking the continuously-typing feature due to the homonym-indication. Therefore, whichever way to achieve the continuously-typing will lead to high input speed. According to character counts for identical documents, the number of English characters is 3.7 times that of Chinese characters. Therefore an input speed of 34 to 68 Chinese characters per minute would be equal to the 125-250 range for English characters (25-50 words.) [WEI85]. Hence 80 Hanzi/min is a very high input speed, 30 to 50 Hanzi/min is acceptable speed and the speed of 15 Hanzi/min or less is too slow.

Criterion 2: Homonym indication

Fast-indication is directly related to a high input speed and it can release users from the burden of homonym-searching so that they can concentrate on writing and thinking. The best way is to let the CWP do the homonym-indication as TianMa does. Besides, some methods such as sorting homonyms according to their usage-frequency and using tones can reduce the size of homonym-list significantly.

To illustrate, the Pinyin method of 中文Talk sorts the homonym-list according to the usage-frequency and achieves a high "hit rate" (about 30 % of characters in a text are the first character on the homonym-list). An example for usefulness of intonation is that the Pinyin syllable "yi" has 164 Chinese characters corresponding to it in the Modern Chinese Dictionary [DICT80]. However, if tones are used then there are only 20 characters with the first tone, 35 characters with the second tone, 20 characters with the third tone and 89 characters with the

Table IV Pinyin Input Method Criteria

No.	Criteria	A	B	C	D
1	Continue-typing and high speed.	typing without indicating the homonym; about 80 Hanzi/min (Hanzi means Chinese character)	typing with reduced indication (e.g. phrase indication, not character indication), 50 Hanzi/min.	stop for indicating each homonym, but "hit rate" is high (e.g. more than 50% of the Hanzi wanted are on the first line), 30 Hanzi/min.	stop for indicating homonym, low hit rate, > 15 Hanzi/min
2	Fast indication.	homonyms sorted by usage-frequence; > 50% hit rate, i.e. first Hanzi is chosen or the Hanzi wanted shown on screen directly (e.g. TianMa CWP).	homonyms sorted by use-frequence; and > 50% in the first line (not necessary the first character).	homonym sorted by use-frequence; and > 50% in the first line if tone is added.	sorted in GB 2312-80 only.
3	Multi-level-compatible Pinyin method, Phrase or glossary abbreviation	full Pinyin and simplified Pinyin compatible; Hanzi/phrase/idiom compatible; part-Pinyin tolerant.	full Pinyin and simplified Pinyin compatible; no phrase and idiom; part-Pinyin tolerant.	no simplified Pinyin; no phrase and idioms; part-Pinyin tolerant.	exact one way of Pinyin.
4	Intonation used properly.	both with and without tones are allowed which improve indication.	---	restricted way, but user feels no inconvenience with his input.	restricted way (either with or without tone but not both); user feels inconvenient.
5	Intelligent Hanzi base or phrase base.	sort Hanzi-base or phrase-base dynamicly for user's typing; capable of producing new Hanzi or phrases.	sort Hanzi-base or phrase-base statically (i.e. manually)	fixed Hanzi-base and phrase base.	no phrase-base.
6	Dialect tolerant.	allow southern China dialect (e.g. Cantonese dialect) allow any slight difference between mandarin and other dialect in China.	allow Cantonese and all slight difference between mandarin and any other dialect.	allow all slight differences between mandarin and any other dialect.	standard Pinyin only.
7	Unknown Hanzi processing.	phonetic way to indicate them, no need to learn component method.	mainly phonetic method with help of other simple way.	other input methods.	looking up dictionary.
8	Error-tolerant auto-correction.	correct wrong characters by context automaticly.	manually indication but automaticly correction.	notice user and prompt him with options.	No way to help user.

fourth tone. Whichever way used to achieve fast indication will improve the hit rate (i.e. the first character is chosen or most characters wanted are on the first line of the homonym list). Criteria 2 is set in this way.

Criterion 3: Multi-level-compatibility

Multi-level-compatibility means to allow users to type in Pinyin in more than one way. The full/simplified Pinyin compatibility means both standard Pinyin and simplified Pinyin are acceptable by a CWP. For example, the Pinyin syllable "hong" may be typed as "hong", i.e. standard Pinyin, or "hs", simplified Pinyin. Those CWP with exact one way of Pinyin like CC-DOS (e.g."hong" must be typed as "hs") violates an important rule, "follow users natural habits".

The full/part compatibility allows users to type only part of the syllable, e.g. "hong" is typed as "hon" or "ho" in case a user is not sure about the syllable. The Character/phrase/idiom compatibility provides a basic way to improve the Pinyin method. Since Chinese text mainly consists of two-character phrases, phrase-input will solve the homonym-search problem (the CWP can easily choose a correct phrase by analyzing context). Similarly, there are many idioms in Chinese text usually taking a four-character form and with an interesting historical story behind each of them. Using a few Pinyin letters to represent an idiom can improve input speed greatly. Criterion 3 is set based on this consideration.

Criterion 4: Intonation

Whether tones should be used or not is questionable. Tones reduce the size of the homonym list, but one more keystroke is needed to indicate the tone. So a CWP should give user freedom to make choice instead of restricting the user to one particular way.

Criterion 5: Hanzi base

Since the Chinese character set and phrase set are very large, no available CWP can provide all characters and phrases needed by all users. An intelligent character-base (dictionary) and phrase-base are essential to CWPs which allow a user to add new character or phrase into the bases, and to sort the character/phrase in the bases in the order which best meets the particular application.

Criterion 6: Dialect tolerant

One of the main problems with standard Pinyin method is its inability to deal with different dialects. Fortunately, people in northern China (which accounts more than half of the

total population) speak the dialects which are very similar to Mandarin, the official Pinyin standard. Middle aged people and young people in southern China know both Mandarin and their local dialect because Mandarin (Pinyin) has been taught since the 1950's. However, their Mandarin is not perfect Pinyin but with slight difference caused by their local dialects. As a result, a CWP should tolerate these differences in some manner. A worst case is the Cantonese dialect which exists in Canton, Hong Kong and most North America Chinese communities. Most Cantonese speakers do not know Mandarin. The only way to solve the problem is to provide Cantonese Pinyin as CCDOS V4.0 Chinese Processing System does [CCDOSV87].

Criterion 7: Unknown Hanzi processing

One major problem with the Pinyin method is the inability to input unknown characters. Fortunately, it is not a serious problem for a CWP because a user usually types what is said instead of copying from available text. A user seldom uses an unknown character. After all, how can people use unknown characters in speach. However, in some cases a CWP user might need to copy something from other sources like books. In this case, the CWP should provide an alternative phonetic way (e.g. an modified Pinyin method) to input the character instead of forcing the user to learn other component methods. In the latter case, the user probably prefers to look up a dictionary than to learn a new component method. CDPinyin (Character Decomposition Pinyin), based on this idea, will be introduced in chapter 8.

Criterion 8: Error-tolerant

Just like the spelling check in English, a way to help users to correct a wrongly chosen character is necessary for a CWP.

5.2.4. Interpretation of Output Criteria (Table V)

Output of Chinese is as important as input although many ignore it. The fact that a Chinese character is an ideograph (i.e.picture based character) makes the task of output of Chinese much more complicated than output of English. CWPs should output characters with a number of styles and sizes, output in a vertical line, from right to left, etc. The following criteria are based on these considerations.

Criterion 1: Character style

A CWP should be capable of outputting characters with various styles such as Song,

Table V Output Criteria

No.	Criteria	A	B	C	D
1	Character-style (font)	provide many styles; good expandability to include more fonts; Chinese/English mixed output; about 10 Chinese calligraphic styles; art lettering; simplified and classical Hanzi.	provide many styles; expandable to some extent, Chinese/English mixed; at least 6 styles, i.e. ImitativeSong, Song, Kai, Bold, Li, and Xing style; simplified and classical Hanzi;	Chinese/English; provide popular styles, i.e. Song ImitativeSong, Bold and Kai; only simplified (or classical) Hanzi provided.	Chinese/English; provide one or two styles only i.e. Song and ImitativeSong; only simplified (or classical) Hanzi provided.
2	Character size	many sizes provided e.g. 16*16, 24*24, 32*32, etc. good expandability.	provide more than 5 sizes, e.g. 16*16, 24*24, 24*16, 32*32, etc.	16*16, 16*24, 24*24.	16*16 only.
3	Special layout e.g. vertical line, right-to-left heading, etc.	support special layout both on screen and printer.	support vertical line output and riht-to-left heading.	deal with special layout in graphical mode. (e.g. the Macintosh)	No such ability.
4	Granularity of style-switching	style can be changed at all three levels,i.e. at the whole text level, at paragraph level, at line level.	style can be changed at two levels, i.e. the character within a line can not be different styles.	style changeable at one level, i.e. the character within a paragraph cannot be different styles.	the whole text can only take one style at a time.
5	Chinese Kerning	even spacing in a line and even the structure of a character.	even spacing in a line only.	No Chinese Kerning.	---
6	WYSIWYG (What you see is what you get)	all styles, sizes, tables, math-symbols printed from printer can be shown on screen.	some (not all) styles, sizes, etc. can be shown on screen.	special symbols are inserted to mark the style (or size) switching.	cannot see any switching on screen (e.g. style or size changes).
7	Fast and efficient output.	fast screen display; background printing; high speed printing.	both screen display and printing are fast, but no background printing and other features.	the speed of screen display and printer is acceptable.	output is slow.
8	Quality of screen and printer	high resolution screen (e.g. Sun station) and high quality printer (e.g. laser printer); other features, (e.g. color printing)	screen with good resolution (e.g. the Macintosh); high quality printer (e.g. laser printer)	no laser printer; screen with good resolution.	ordinary devices (e.g. dot matrix printer only).

Kai, Li, Bold, etc. The author did some simple statistical analysis: four newspapers and magazines are chosen. Among them, two are published in mainland and other two in Taiwan. Five samples are randomly chosen for each of them and the average number of character for a style is calculated. Although the results may not be accurate, they tell the basic fact. Fig. 5.1 and 5.2 shows the usage frequency of popular character styles for some newspapers and magazines published in China and Taiwan. Table VI shows the number of characters for each style. Table VII shows the character styles provided by 11 CWPs which indicates that the importance of multi-style output has not been realized by many because most CWP provide one or only a few styles and only two of them provide Kai style, a popular style used in the Chinese publishing industry. Besides, a CWP should not only provide multi-style output but also allow users to add their own in cases where special styles (e.g. Zhuan style) are required.

Criterion 2: Character size

Currently, most CWPs provide more than one character size (e.g. 16*16 and 24*24). Considering the style of Imitative Song (16*24 or 24*16) at least 4 or 5 sizes are required.

Criterion 3: Special layout

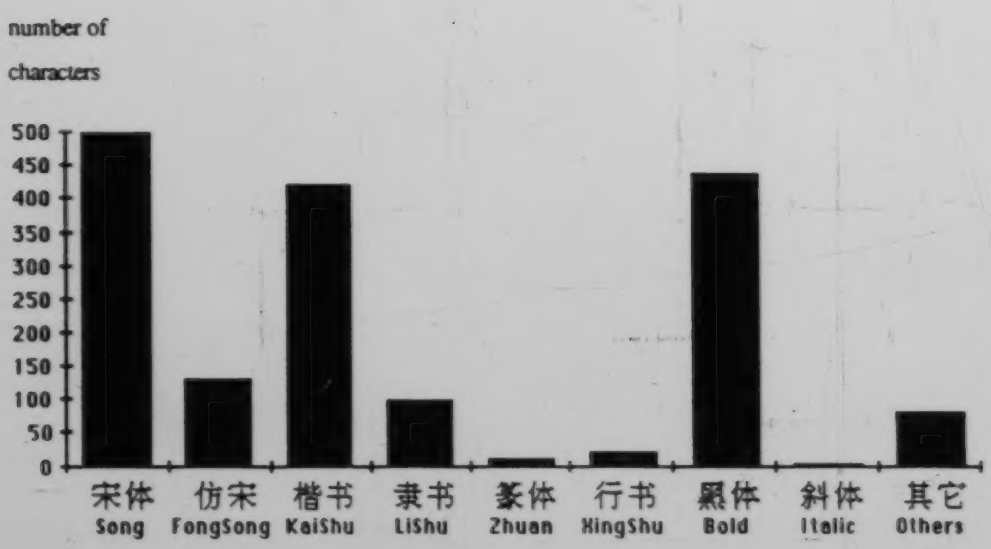
Some special layouts such as vertical and right-left are not popular in China, but very popular in Taiwan, Hong Kong and overseas Chinese communities. Therefore a CWP should be capable of generating such layouts. While graphics software can generate these by "drawing" the text, this is not the way a CWP should work.

Criterion 4: Style-switching

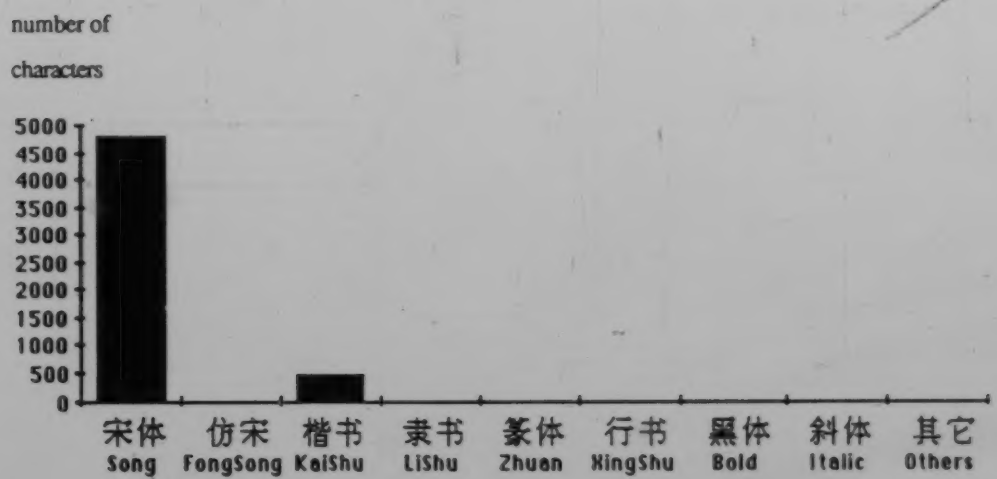
There are 3 levels at which the character styles may be changed, that is, the character style in a line can be different, the character style in a paragraph can be different (but not in a line), and the character style in an article can be different (i.e. can only change the whole paragraph). Obviously the smaller the granularity, the more powerful the CWP is.

Criterion 5: Kerning

Kerning is the process of adjusting the space between specific characters to create more even letter spacing. Kerning is not a big problem for Chinese text because Chinese characters are all square. If the definition of kerning is expanded to include keeping a balanced character structure, then it becomes very important for CWPs, which means that the characters in the font library should have a balanced structure and facilities should be provided to help users design their own fonts and characters.

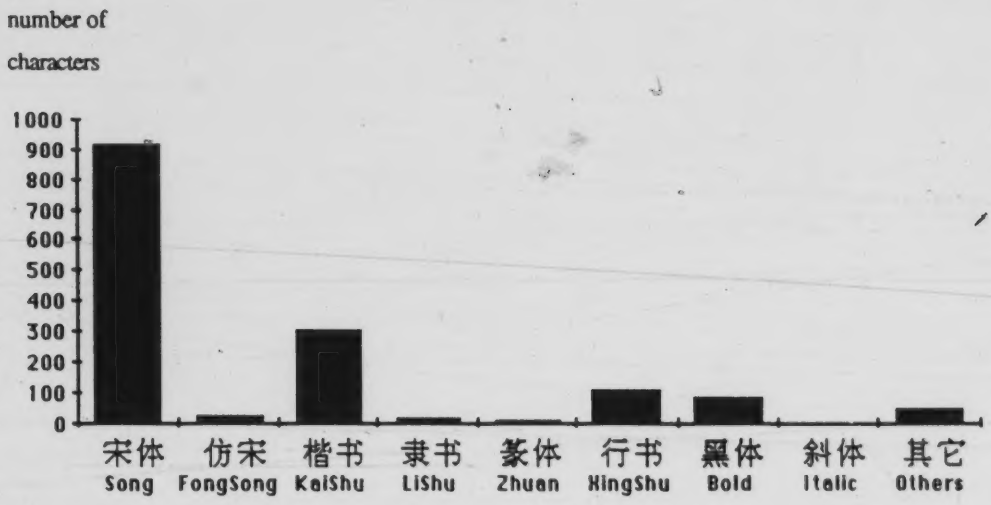


(a) 人民日报 (标题) People's Daily (Headline)

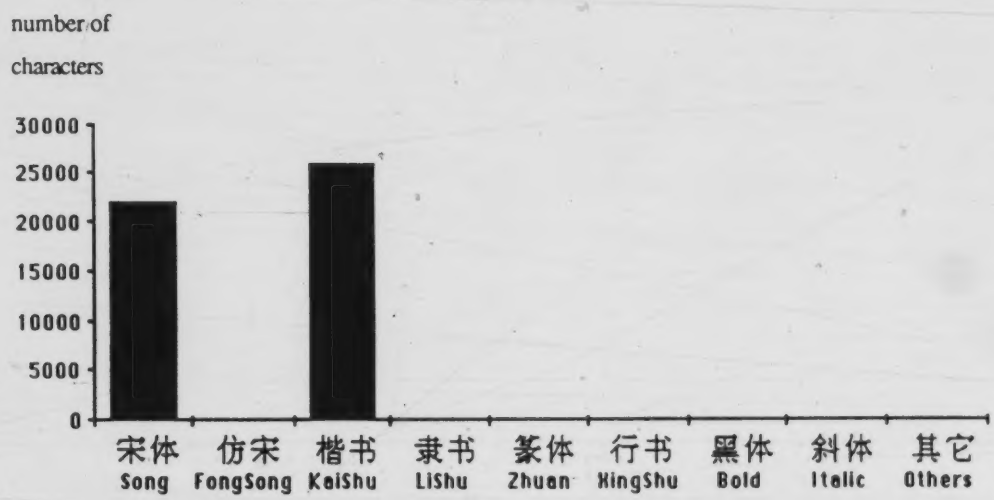


(b) 人民日报 (文章) People's Daily (Articles)

Fig. 5.1. The Usage-frequency of Character Styles on People's Daily (China) [PD88]



(a) 海外学人(标题) Chinese Scholar Abroad (Headline)



(b) 海外学人(文章) Chinese Scholar Abroad (Articles)

Fig. 5.2. The Usage-frequency of Character Styles on Chinese Scholar Abroad (Taiwan) [CSA88].

Table VI. The Usage Frequency (# of Characters) of Popular Character Styles for Some Newspapers and Magazines

Character Style	人民日报 People's Daily		中央日报 Central Daily		神洲学人 Chinese Scholar		海外学人 Chinese Abroad	
	标题 Heading	文章 Articles	标题 Heading	文章 Articles	标题 Heading	文章 Articles	标题 Heading	文章 Articles
宋 Song Style	519	46200	568	48000	173	14020	920	22000
仿宋 Fang Song	145		104		204	13660	25	
楷书 Kai Shu	432	1820	180		20	13588	302	26000
隶书 Li Shu	109		32					
魏碑 Wei Bei					10			
篆 Zhuan	5						5	
行书 Xing Shu	13		6		49		108	
黑体 Bold	430	930	460		132	720	90	
斜体 Italic	8				28			
围边 Hollow	54		12					
Others	30		30		6		45	
Total	48000		48000		84480		48000	

Table VII. The Character Styles Provided by Some CWPs and Chinese System

CWP	宋 Song	仿宋 FangSong	黑体 Bold	楷书 KaiShu	隶书 LiShu	行书 XingShu	魏碑 WeiBei	篆 Zhuan
System C (Wang US-100)	✓	✓		✓				
中文Talk (Macintosh)	✓							
天马 TianMa (IBM-PC)	✓	✓						
凯兴 KaiHin (Macintosh)	✓	✓						
长城GreatWall (GW 0520-CH)	✓	✓	✓	✓				
CC-DOS V4.0 (GW 0520-ABC)	✓	✓						
汉字YS-WORD (ZD-3100 Image Computer & PC)	✓	✓						
HZ-PC汉字系统 (西子Chinese Computer)	✓	✓						
ICIP-2 (GW 0520-CH)	✓	✓	✓					
MPC-DOS V3.1A (IBM-PC)	✓	✓						
QPC-DOS (IBM-PC)	✓	✓						
Eagle PC-Plus-XL 汉字微电脑	✓							

Criterion 6: WYSIWYG

"What you see is what you get" means that everything output is shown on the screen. This is an important feature of word processing and CWP is no exception. By "everything" it means the character styles and sizes, tables, mathematical equation, graphs, etc. Usually, graphics-based computers like the Macintosh and Sun have the feature WYSIWYG. Many CWPs do not have this feature and special symbols are inserted into the text to mark the layout (style, table, etc.). Then, the printer-controller recognizes the symbols to generate the appropriate layout. The main drawback of this approach is that user has no clear idea of what the output looks like until the hard copy is produced. Usually, it is necessary to edit and print the text many times to get a satisfactory layout. Therefore the more the user can see on the screen, the easier it is to edit the document, and the more efficient a CWP will be.

Criterion 7: Speed

Fast output is a common requirement for a word processor and it is even more important for a CWP because of the ideographic nature of the language (note: displaying a graph is much slower than displaying text). In addition, other facilities, such as background printing, may be invoked to improve output efficiency.

Criterion 8: Output devices

The quality of output devices are crucial for CWPs due to the fact that an English letter takes an 8*8 bit map on the screen while a Chinese character takes at least a 16*16 bit map. Actually, the characters acceptable by most users is a 24 * 24 bit map. Considering the character quality and the output speed a laser printer is necessary.

5.2.5. Interpretation of Editing/Formatting Criteria (Table VIII)

The editing/formatting aspects of Chinese word processors are not very different from English word processors. In some aspects the former is even simpler than the latter. For example, word wrapping is an easy matter for Chinese characters because of the square shape and equal size while English words have different lengths. For the same reason a CWP does not need the soft-hyphen feature of an English word processor. As a result the editing/formatting criteria is simpler than the others and may not need four levels for each criterion.

Table VIII. Editing/Formatting Criteria

No.	Criteria	A	B	C	D
1	Friendly user interface.	powerful window system including multi-window, fast window operation, zoom, move, etc. pull-down (or pop) menus, mice, icon, pointer, etc. easy-operation like cut, paste, etc.	a good window system with limited number of windows and acceptable speed of window operation. part of other facilities such as icon, menu, etc.	slow and limited window operation.	Lack the way to help user interact with CWP.
2	Headers and footers, etc.	all functions e.g. headers, footers, page number, foot notes, time/date stamping, etc.	provide part of these functions.	—	no such functions provided.
3	Page formatting.	all functions e.g. multi-columns, copy ruler or paragraph format, style sheet, table of content, line spacing, etc.	provide general functions only, e.g. set margin, line spacing, etc.	—	no such functions provided.
4	Type control.	all type control functions e.g. tab, decimal tab, superscript, subscript, non-printing character showing, etc.	provide part of type control functions e.g. tab, decimal tab, etc.	—	no such functions provided.
5	Graphics	deal with graphics and text in a integrated way, e.g. import graphics, direct import graphics, wrap text around graphs, and put graphs in the same line with text.	provide part of such functions e.g. combining text and graph in the same document.	—	no such functions provided.
6	Utilities	all utilities e.g. mail merge, page preview, character/phrase count, on-line-help, file conversion (e.g. Pinyin/Hanzi conversion, character style conversion, etc.) font map display.	provide commonly used utilities only, e.g. on-line-help, style conversion, etc.	—	no such functions provided.

Criterion 1: Friendly user interface

The main advantage of a word processor over a typewriter is its editing and formatting functions and since users need to interact with the word processor frequently to edit or format documents, a friendly user interface is indispensable. Besides, a word processor provides editing operations such as inserting, deleting, setting margins, etc.; however, not all of them provide friendly user interfaces. In fact graphics-based computers such as the Macintosh and Sun obtain a powerful user interface which makes them very successful in word processing. Based on the state of the art of microcomputers, a set of powerful window operations is necessary (together with other facilities such as menus, icons, mice, etc.). Criterion 1 is set according to this consideration.

Criterion 2: Headers and footers

Headers, footers, page number, etc. are widely used in English word processors and are also used in CWPs.

Criterion 3: Page fomatting

As word processors are becoming powerful, more and more page formatting functions are added. A powerful CWP may include some advanced features such as multi-columns and style sheets. A common CWP should at least provide the page formatting functions such as set-margin, centering, line spacing, etc.

Criterion 4: Type control

Type control for various characters is necessary especially for scientific writing. A CWP should at least provide controls such as tab, decimal tab, superscript, and subscript.

Criterion 5: Graphics

To generate more complicated documents, a graphics capability is a necessary part of the word processor. A CWP should be capable of processing both text and graphics.

Criterion 6: Utility

Word processors (including CWPs) are very popular software packages which change rapidly. This criterion takes more advanced features of CWPs into consideration.

5.3. WEIGHT ASSIGNMENT

To evaluate a CWP in a quantitative way, a hundred-point scale is adapted that represents the weights attached to the criteria in Section 2. The primary consideration of assigning weights to the criteria is the following:

(1) The weight assignment should match the objective evaluation of existing CWPs. However, since there are not many CWPs currently available the weight assignment may not be accurate. When more and more CWPs exist they will be more accurate and objective.

(2) More points should be assigned to the criteria which reflect unique features of CWPs.

(3) The weight assignment should reflect the different extent to which the criterion affects the quality of the CWP.

(4) The weight assignment below shows the author's opinion and it is certainly not standard. Any user is free to set weights to the criteria to meet any unique requirement (e.g. input method, output quality, or editing function, etc.). However, the weight assignment should be as objective as possible so that a good CWP can be recognized.

The distribution of points is:

(1) A total of 100 points are assigned to Input method, Output method, and Editing/ formatting criteria. Among them the Input method and the Output method are each assigned 40 points and Editing/formatting 20 points. The reason is that the input and output parts of a CWP show the unique features of Chinese processing and have a great influence on the performance of the CWP. While the editing/formatting part of a CWP is not very different from an English word processor.

(2) Both Input and Output are assigned 40 points which shows that they are equally important. Many pay more attention to the input method than the output method. If there is a breakthrough in the input method then Chinese processing will be successful. Actually this is not necessarily true. A CWP with a poor output function (e.g. the screen and the printer are slow and an ugly character-layout) cannot be accepted by users. Moreover, Chinese output is not an easy task as people think. For example, designing an output facility which provides enough character styles and sizes and supports the layout of vertical line and right-to-left-heading, and has reasonable time and space costs, is certainly not an easy task.

(3) There are four levels for each criterion, i.e. A, B, C, and D which represent 100 %, 75 %, 50 %, and 25 % respectively. Suppose a criterion's weight is 10 points, then level A is 10 * 100% = 10, level B is 10 * 75% = 7.5, level C is 10 * 50% = 5, and level D is 10 * 25% = 2.5.

Moreover, there are three sublevels attached to each level, i.e. upper, middle, and lower sublevel respectively. The middle sublevel is the level itself, the upper sublevel is the level plus 1/3 of the interval and the lower level minus 1/3 of the interval. Usually the level A to D are used and the sublevels are needed only when a user wants more accurate results to separate two similar CWPs.

The following are the detailed assignment of weight to each table.

Table IX. The Weight Assignment to Input Method Criteria: (40 points)

No.	Criteria	Weight (points)
1	According with user's natural habits	10
2	Easily-remembered coding rule and short learning time	7
3	Duplicate code rate and number of strokes per character	2
4	Input-method-interface	4
5	Input speed and touch-typing	8
6	Size of keyboard and extra hardware needed	4
7	Facility and higher degree of intelligence	5

Justification:

The weight for criteria 1 and 3 show that more attention is paid to the complete performance of a CWP instead of an individual criterion like duplicate code because a good result for individual criterion does not necessarily imply a good CWP. For example, to reduce the duplicate code rate a software developer has to design a

complicated coding rule which is hard for users to handle. A CWP is used as a whole facility and does not always put much attention to an individual criterion [JIANG85]. Among the others, criterion 2 and 5 have more points because they imply that a user is released from the burden of coding and may concentrate on writing. Criterion 7 gets more points than criteria 4 and 6 because the more intelligence a CWP has, the easier it can be used. A CWP developer should think more about it than items such as the number of input methods.

Currently most CWPs use a standard keyboard and try to eliminate extra hardware, which makes the Criterion 6 less important. Therefore the criterion is not given many points.

Table X. The Weight Assignment to Pinyin Input Criteria: (40 points)

No.	Criteria	Weight (points)
1	Continuous-typing and high speed	10
2	Fast indication	4
3	Multi-level-compatible Pinyin, phrase abbreviation	3
4	Intonation used properly	3
5	Intelligent Hanzi base and phrase base	7
6	Dialect tolerant	7
7	Unknown Hanzi processing	3
8	Error-tolerant and auto-correction	3

Justification:

Criterion 1 is given the biggest weight because it is the main problem for most Pinyin methods in

the sense that typing is often interrupted by homonym searching. This problem is so crucial to Pinyin that any breakthrough in continuous typing will lead to a good Pinyin input method (e.g. TianMa [TianMa86] and ShuangPin [Liu87]).

Criterion 6 is given a bigger weight because dialect is a major problem with most Pinyin methods. Any breakthrough in dialect problem will lead to a good Pinyin input method (to the author's knowledge, only CCDOS V4.0 supports Cantonese Pinyin besides standard Pinyin). Criterion 5 is also more important than the rest because it is phrase (and idiom) input that makes continuously-typing and fast indication possible. Intelligent Hanzi and phrase bases are indispensable for phrase input.

Among the other 5 criteria the 2nd obtains one more point because it is closely related to criterion 1(the most important aspect). All the rest obtain the same number of points since they are all auxiliary ways to help solve the major Pinyin problems. Although the 7th criterion is also a major Pinyin problem, it is not a serious problem for CWPs because users seldom speak or write characters unknown to them.

Table XI. The weight assignment to Output Criteria: (40 points)

No.	Criteria	Weight (points)
1	Character style (font)	7
2	Character sizes	3
3	Special layout (e.g. vertical line, right to left heading)	2
4	Granularity of style-switching	3
5	Chinese kerning	2
6	WYSIWYG (what you see is what you get)	10
7	Fast and efficient output	8
8	The quality of screen and printer	5

Justification:

Among the 8 criteria for Output the 6th called "what you see is what you get" has a great influence on whether a CWP is user-friendly because this feature allows a user to know exactly what the document will look like before it is actually printed out. Most word processors (not only CWP) do not provide this important function. Users not using a WYSIWYG word processor have the painful experience of going back and forth between the hard copy and screen editing for many times before a satisfactory document is obtained.

Criterion 7, fast and efficient output, is a key issue of output due to the fact that output devices have to work in graphics mode because of the ideographic nature of Chinese. This usually slows down the output speed so that a CWP is not easily accepted by many users.

Criterion 1, character style, is another key issue of Chinese output. Although the importance of character style has been noticed, it will become clearer when more and more CWPs exist. At that time, users may not accept a personal letter printed out by a CWP with Song style characters (it looks like a government document instead of a letter to his wife!). The other reason for this criterion to get more points is that implementing a multi-style is not easy because of the storage cost.

The next criterion with a bigger weight is the 8th, output devices, because high quality Chinese output is impossible without high quality devices.

All the rest are not crucial issues of an output method, hence smaller weights (2 or 3 points) are assigned to them. For example, multi-size is not hard to implement and most CWPs have this feature. Also most Chinese people do not use right to left headings.

Table XII. The Weight Assignment to Editing/Formatting Criteria: (20 points)

No.	Criteria	Weight (points)
1	Friendly user interface	5.0
2	Headers and footers, etc.	2.5
3	Page formatting	2.5
4	Type control	2.5
5	Graphics	2.5
6	Utilities (facility)	5.0

Justification:

The editing/formatting aspect of CWP is very similar to that of an English word processor so only 20 points are given to it. Based on the same ideas mentioned before, (i.e. judging a CWP as a whole instead of emphasizing individual elements), Criteria 1 and 6 are given the biggest weight and all others get the same weight because they are common features of a word processor (nothing unique to a CWP) with similar importance.

5. 4. Experimental Test of the Available CWP

Although the proposed evaluation criteria is as detailed as possible it cannot include every aspect for all CWP, therefore experimental testing is necessary. Besides, it can show the validity of the proposed evaluation criteria (note that the results of evaluating the available CWPs by three methods are given in Appendix A).

5.4.1. Description of the Experiments

The following experiments are designed for the Pinyin input method only because:
(1) Pinyin is the only input method available for some CWP (e.g. TianMa CWP).
(2) Pinyin is believed to become the first national standard input method and it will be discussed in more detail in Chapter 8.

In experiment 1 (Fig. 5.3), a paragraph taken from practical writing (a piece of news from a newspaper) is used to test the input/output speed of the available CWP. Because it is practical writing most characters are high-frequency-character (i.e. popular in daily life) therefore the hit-rate is going to be high for the phrase base of TianMa. Table XIII shows the results.

Table XIII. The Results of Experiment 1 (News with 291 characters)
(unit: minute)

	KaiHin	中文Talk	TainMa
Input Speed	49	39	17.2
Output Speed	1.8	1.8	5.3 (big) 2.1 (small)

用计算机处理的信息，包括数据，文字，图形，语音等，其中主要的是文字信息．我国是个多民族的国家，大多数人民属于汉族，因此我国使用的主要是汉字．新中国成立以后，承认全国各民族一律平等，一反对民族沙文主义和狭隘民族主义，因此有必要区分"汉字"和"中文"的提法，不能把它们混同起来．中文应包括少数民族的文字．我国要推广计算机应用，必须使各用民族的文字都能在计算机上实现输入和输出．这些民族的文字中，只有汉字是一种象形文字，其他的都是拼音文字．拼音文字的字符最多不超过数十种，而现代我国使用的汉字仅常用部分就有六，七千字，而且字型复杂，这就使计算机实现汉字信息处理存在较多困难．如果解决了汉字信息处理的难题，对于我们其他民族的文字处理就比较容易解决．

Fig. 5.3 A Piece from a Newspaper.

将进酒　　李白

君不见黄河之水天上来，奔流到海不复回．
君不见高堂明镜悲白发，朝如青丝暮成雪．
人生得意须尽欢，　莫使金樽空对月．
天生我材必有用，　千金散尽还复来．
烹羊宰牛且为乐，　会须一饮三百杯．
岑夫子，　丹丘生，　将进酒，杯莫停．
与君歌一曲，　请君为我倾耳听．
钟鼓馔玉不足贵，　但愿长醉不复醒．
古来圣贤皆寂寞，　惟有饮者留其名．
陈王昔时宴平乐，　斗酒十千恣欢谑．
主人何为言少钱，　径须沽取对君酌．
五花马，千金裘，　呼儿将出换美酒，
与尔同销万古愁．

Fig. 5.4 A Classical Poem

In experiment 2, a classical poem, shown in Fig. 5.4, is used to test the input/output speed of the CWP. Because a classical poem tends to use low-frequency-characters and phrases (i.e. not always popular for ordinary people) the hit-rate is going to be low for the phrase base of TianMa (actually a classical poem is almost the worst case for TianMa). On the other hand, it will not affect KaiHin and 中文 Talk because they select characters (not phrase) from the homonym list. The following result shows it.

Table XIV. The Results of Experiment 2 (Poem with 176 characters)
(unit: minute)

	KaiHin	中文Talk	TainMa
Input Speed	41	31	28.3
Output Speed	1.1	1.1	3.4 (big) 1.2 (small)

5.4.2. Analysis of Above Experiments

中文 Talk:

formula for input time:

Total Time = Tt + (To) + Ts

where Tt = typing time
(To) = time for tone indication (optional)
Ts = time for searching the homonym list
(character is chosen by mouse)

News:

hit rate = 73 % (i.e. the first character chosen)

tone: 25 % of characters in the paragraph need tone to bring them to the first line of the homonym list.

Poem:

hit rate = 55 % (i.e. the first character chosen)

tone: 25 % of characters in the paragraph need tone to bring them to the first line of the homonym list.

Comment:

Using a mouse to indicate homonym is much slower than using number key (KaiHin).

The time cost for searching a character in the homonym list is:

time = catch_mouse + locate_character + click + back_ to_keyboard

Note: 中文 Talk allows to indicate a homonym by arrow key. However, in most cases, using arrow key is still slower than using number key.

Arranging homonym list in use-frequency is successful which leads to high hit rate.

As a result, 中文 Talk is more efficient than KaiHin although mouse-indication is slow.

KaiHin:

formula for input time:

Total Time = Tt + Ts

where Tt = typing time

Ts = time for searching homonym list

(character is chosen by number key)

News:

hit rate = 20 %

Poem:

hit rate = 15 %

Comment:

The hit rate is low because the homonym list is arranged according to GB 2321-80 which leads to slow input speed.

TianMa:

formula of input time:

Total Time = Tt + (Tc) + Te

where Tt = typing time

(Tc) = time for Pinyin/Chinese conversion (if automatic-conversion is used Tc= 0)

Te = error correction (in case wrong character or phrase is chosen)

News:

Tt + (Tc) + Te

= 12 min + 14 sec + 5 min = 17 min 14 sec

error rate: 35 / 291 = 11 %

Poem:

Tt + (Tc) + Te

= 7 min + 80 sec + 20 min = 28 min 20 sec

error rate: 107 / 176 = 60 %

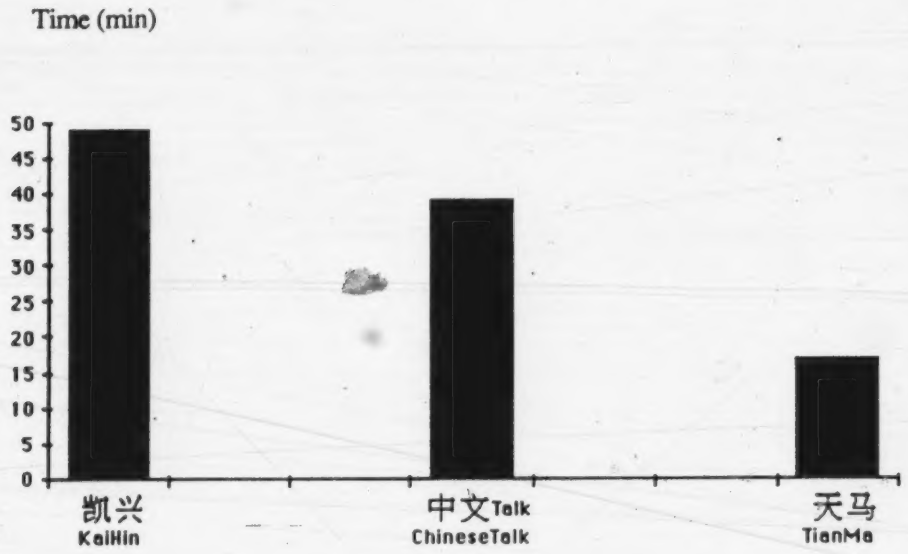
Comment:

TianMa is much better than 中文 Talk and KaiHin (Fig. 5.5). Even in the worst case (for the classical poem the error rate is 60 %), TianMa is still better than others because of:

(1) smooth and continuous typing.

(2) many error corrections are for phrase-correction which is still faster than character by character indicating of homonyms.

Experiment 1: News (291 characters)



Experiment 2: Poem (179 characters)

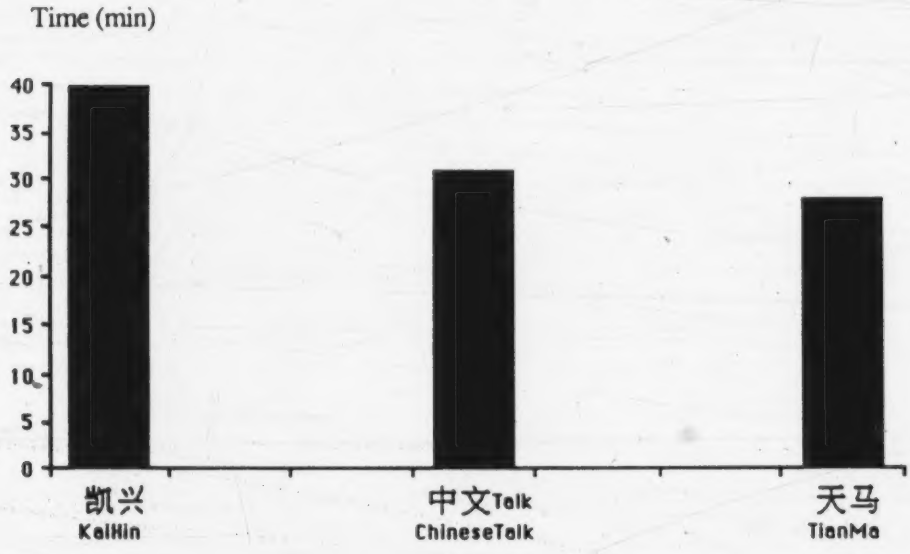


Fig. 5.5 The Comparison of Pinyin Input Speed

Comment:

The experimental test results are consistent to that of the proposed evaluation criteria, that is, TianMa CWP is better than the others and its scheme of continuous-Pinyin-typing and a phrase-base is very successful. Graphics-based computers such as the Macintosh provide a good foundation for CWP development. A CWP under the support of 中文 Talk can be a very powerful CWP because it will make full use of the features provided by the graphics-based computer. Unfortunately, many Chinese computer scientists and software developers have not taken advantage of it. People in China have not even noticed the importance of CWPs. The author believes that the situation will change with the development of office automation. More and more CWPs coming into being will also help to solve the problem.

5.5. Conclusion

In this chapter detailed criteria for the evaluation of Chinese word processors (CWP) are established, which attempts to evaluate CWPs in their entirety. The results by applying the criteria to three existing CWPs (i.e. KaiHin, TianMa and 中文Talk) are consistent with experience and the results of experimental tests. The method is useful for both evaluating existing CWPs and designing a new CWP. The results show that the TianMa CWP solved the homonym-indication problem of Pinyin and made a practical Pinyin input method possible. The results also indicate that graphics-based computers such as the Macintosh provide a good foundation for CWP development. In addition, Character Decomposition Pinyin, an improvement to TianMa CWP, is introduced in Chapter 8.

Chapter 6
CHINESE ELECTRONIC PUBLISHING (CEP)

Chinese electronic publishing (CEP) is another important application of Chinese computing due to its theoretical and practical significance. Theoretically, the most outstanding features of Chinese computing are reflected as its input and output aspects, and CEP shows output features to the full. Practically, replacing manual-typesetting by CEP can improve the publishing efficiency greatly. This can also be seen by the fact: the first Chinese computing research project in China was established in August 1974 and among three research topics, the Chinese electronic publishing took the highest ranking (the other two were a Chinese information retrieval system and a Chinese communication system) [GUO85].

High printing quality is of primary importance in book, magazine and newspaper publishing, the printing resolution requirements should be well over 600 dots/inch, and probably closer to 2400 dots/inch. More than ten typefaces are frequently used in Chinese publications, each typeface should contain more than 7,000 Chinese characters with more than ten sizes. This implies that the total font storage required for these Chinese printing systems is over two billion bytes. Therefore, an efficient representation and generation of high resolution Chinese characters becomes a key problem in Chinese electronic publishing systems.

In this chapter, the problems and technical difficulties of Chinese electronic publishing is described, a successful publishing system, the Hua-Guang system by Beijing University and other organizations, is introduced in depth, and a general survey of other available Chinese publishing systems is also presented.

6.1. Publishing Problems

Currently laser typesetters are widely used in electronic publishing and have the advantages of generating high quality characters, fast output speed, good adaptability, capable of outputting black/white image and photos, etc. Its most attractive feature is so called no-film plate making. The disadvantage is that the control of laser scanning is not very flexible and the size of the dot by a laser beam can not be changed instantly. Because of the second disadvantage, characters with different sizes must be represented by different dot matrices. Table XV shows the relationship between different character sizes and their dot matrices.

Table XV. The Relationship between Chinese Publishing Character Sizes and Their Dot Matrices

Size-Name	Point (Type)	Dot Matrix
Seven	6-point	62 * 62
Small Six	7-point	72 * 72
Six	7.875-point	82 * 82
Small Five	9-point	92 * 92
Five	10.5-point	108 * 108
Small Four	12-point	124 * 124
Four	14-point	144 * 144
Three	15.75-point	162 * 162
Small Two	18-point	184 * 184
Two	21-point	216 * 216
One	28-point	288 * 288
Small First	31.5-point	324 * 324
First	35-point	370 * 370
Small Special	42-point	432 * 432
Special	49-point	494 * 494
Special Big	56-point	576 * 576

(Note: the data on the table is based on 29.2 scan-lines / mm. A **point (type)** is a unit of measure in publishing industry and 1 point = 0.35 mm).

There are many styles of publishing Chinese characters such as: Song (宋) for books, Song for newspapers, Heading Song, Imitative Song (仿宋), Regular (or Kai 楷), Bold (黑体), Narrow Song (窄宋), Wide Song (扁宋), Narrow Bold, Wide Bold, Li (隶书), etc.. There are more than 7,000 characters for each style. Therefore the total number of Chinese characters for publishing purposes (including different sizes and styles) is more than

650,000 which requires the storage space of more than $2 * 10^9$ (i.e. 2 GB). Obviously, this is a key problem in CWP and how to generate high quality Chinese characters from a compressed representation of the characters is the central issue of this chapter.

In section 6.2 two compression methods, run-length-code and radical composition method, are introduced.

6.2. Run-Length Code and Radical Composition Methods

6.2.1. Run-Length Code

This is a simple data compression method first used for a CRT typesetter (called Digiset) of Hell Company in West Germany in 1965. Instead of recording all bits in a dot matrix, the method records the length of each white and black segments of the dot matrix. The format of a row information is:

{row flag}	{# of repetitions}	{white}	{black}	{white}	{black} ...	{white}	{black}
*	n	W_1	B_1	W_2	B_2	W_m	B_m

when there exist n adjacent rows with the same white and black segments, one row of information is enough to represent these rows (the number of repetition = n). In the above format, the row flag and the number of repetitions take one byte, a white segment (e.g. W1) takes one byte and a black segment (e.g. B1) takes one byte (Fig. 6.1).

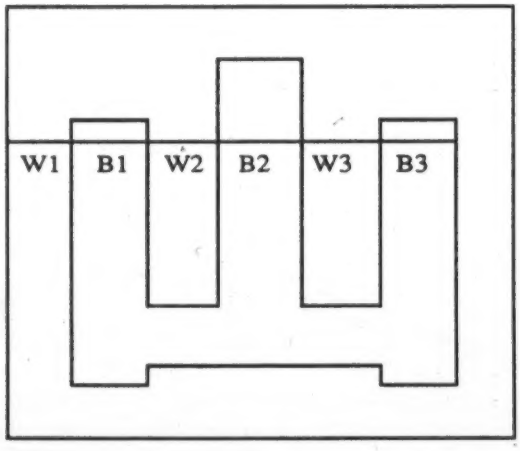


Fig. 6.1 Data compression by recording black and white segments [ZHANG84]

For example, a character with dot matrix 256 * 256 contains the average of 3 black segments per row, hence there are 7 bytes per row and the total spaced needed by run-length code is 256 * 7 bytes. While originally, the dot matrix needs 256 * 32 bytes.

Although a run length code is very simple, it can still recover the original dot matrix accurately. In addition, the hardware for recovery is not complicated. However, the compression ratio is too low and it is not easy to generate other sizes based on one character size (i.e. different sizes for the same character have to occupy different storage space). As a result, a 80 MB disk can only provide 10 to 20K characters (dot matrix).

6.2.2. Radical Composition

This is a data compression method based on the fact that a Chinese character is composed of radicals. Instead of storing tens of thousands of characters in the memory of a computer, it stores hundreds of radicals. All characters can be composed by changing the shape and position of the radical, i.e. by enlarging, shrinking, lengthening, widening the radicals and locating them at the proper positions of a square plate to form a character.

For example,

the character "新" is composed of three radicals "立" "木" "斤" in the following way:

- reduce the size of the radical "立" by 1/2 in both X and Y directions and put it at the upper left corner of the square plate,
- reduce the size of the radical "木" by 1/2 in both X and Y directions and put it at the bottom left corner of the square plate,
- reduce the size of the radical "斤" by 1/2 in the X direction and put it in the right part of the square plate.

Because the number of radicals is much less than the number of characters the method can obtain a high compression ratio. However, a serious problem with the method is the poor quality of the composed characters, which is caused by the following two reasons:

(1) The structure of published Chinese characters is very rigorous and the structure of a character by radical-composition is often different from that of the standard publishing character. Statistics on a simple radical "艹" shows that the radical is very changeable which is influenced by other parts of the characters. So generating all the characters with "艹" by only a few variations of "艹" will lead to poor quality characters.

(2) The thickness (width) of strokes of a composition character is often out of proportion.

For example, the horizontal lines in the characters "立" and "新" are similar in width (i.e. the proportion is about 3:4 or 4:5). The composition character "新" obtains much thinner horizontal line than that of "立" (i.e. about 1:2 or even 2:5) which m kes the character "新" look ugly.

Due to above reasons, radical composition method is not widely used in Chinese publishing systems.

As previously discussed, it is a difficult problem to obtain high quality character with a high compression ratio. (It is now easier to understand that output of Chinese is not an easier task than the input of Chinese as many people thought in the past).

In 1979 Beijing University and other organizations developed a Chinese editing-publishing system which obtained high quality characters and a high compression ratio (called Hua-Guang system). In 1983, the system was updated by VLSI and microprocessors. The following section will describe the system in depth.

6.3. A Technique for Generating High Quality Characters with High Compression Ratio

The method used by Beijing University has the following features:

- Basic strokes (not radicals) are classified into regular and irregular strokes in order to describe them in detail by compression representation and to ease the character enlarging and shrinking;
- Two special custom microprocessors are used to transform the compressed representation of Chinese characters into dot matrices, which solves the difficulty introduced by laser typesetter (see 6.1);
- There are two different styles of composition software for Chinese typesetting, batch mode and interaction mode. The latter is very useful for a newspaper page makeup system.

6.3.1. Regular and Irregular Strokes

In this system the strokes of Chinese characters are classified into regular strokes including

both horizontal, vertical and turn strokes composed mainly of strait segments and irregular strokes in the form of curves. For regular strokes, a compressed 24 to 48 bit representation is used to describe the beginning coordinates, length, width, starting feature, ending feature, turning feature and slant (Fig. 6.2).

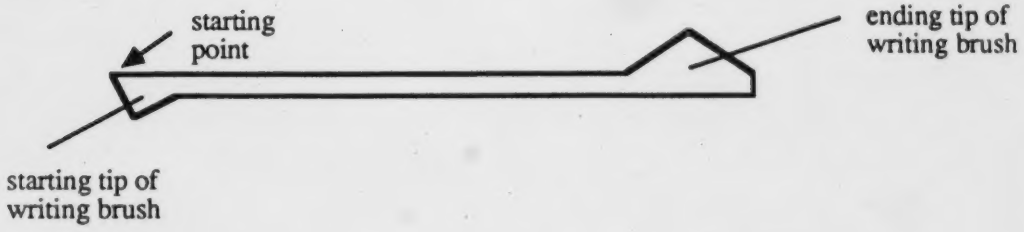


Fig. 6.2 The horizontal stroke of Song style character (regular stroke) [GUO85]

The point of recording basic strokes instead of radicals (see Section 6.2.2) is that the shape of a stroke does not vary too much, so the strokes can be described by compressed representation easily and furthermore a character can be accurately composed by basic strokes. Whereas radicals do vary greatly, which affects the quality of characters. Figure 6.2 shows a horizontal stroke of Song style and there are only 7 varieties of such strokes (i.e. they have different starting and ending features). Hence, 3 bits are enough to represent the starting and ending features of the stroke (i.e. numbers 0 to 6), plus other bits for starting point (x, y coordinates), length, width, etc.. The horizontal stroke can be accurately represented by 3 to 4 bytes without any distortion.

Note that more than half of the strokes of Chinese characters are regular and the accurate description of the strokes (described above) plays an very important role in increasing the compression ratio and keeping high quality when enlarging or shrinking the characters (the width and contrast of basic strokes have a great influence on character quality).

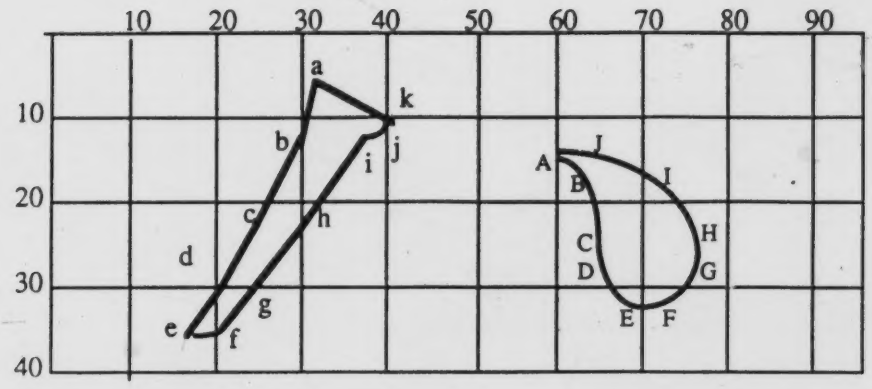


Fig. 6.3 Two irregular strokes [GUO85]

For irregular strokes, a series of vectors are used to approximate the curved contour. Fig. 6.3 shows two irregular strokes represented by series of vectors {ab, bc, cd, de, ef, fg, gh, hi, ij, jk, ka} and {AB, BC, CD, DE, EF, FG, GH, HI, IJ, JA}. The numerical representation for the second one is:

+3, +3; +2, +5; +1, +5; +3, +2; +4, -1; +3, -3; +0, -3; -3, -3; -7, -4

Note that a 5-point character is the basic character size in the Chinese publishing industry which has a dot matrix of 108 * 108. In a computer system only 5-point characters are stored and all other sizes are generated by enlarging and shrinking 5-point characters (by hardware). Therefore every character can be seen as an image on a 96 * 96 dot matrix (resulting from removing the interval among adjacent characters from 108 * 108 dot matrix) and Fig. 6.3 displays such examples.

A 5-point character with its regular and irregular strokes represented by the above compressed method needs an average of 100 bytes of storage without distortion.

6.3.2. Chinese Typeface Recovery

Chinese typeface recovery transforms the compressed representation of the characters into dot matrices for output, and involves three steps:

(1) obtain a series of vectors by analyzing the compressed representation of regular and irregular strokes, enlarge or shrink the series of vectors to obtain the characters with required sizes.

(2) transform the vectors to the closest pixel position of the dot matrix, mark the pixel position with a 2-bit flag to form a flag matrix. The 2-bit flag stands for the following four cases:

00 --- empty space (or white segment);

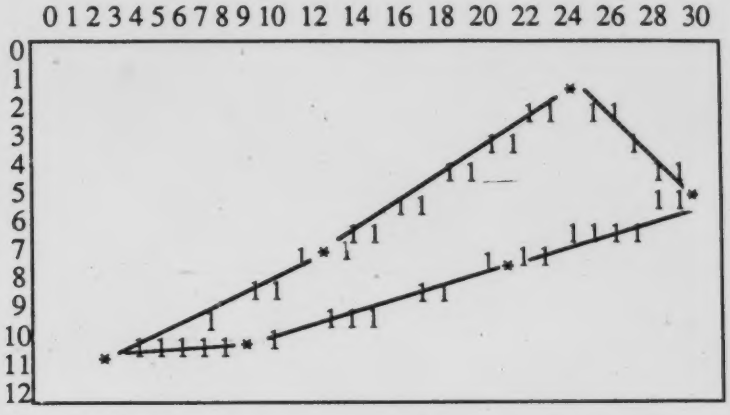
01 --- starting point of a black segment;

10 --- ending point of a black segment;

11 --- black point.

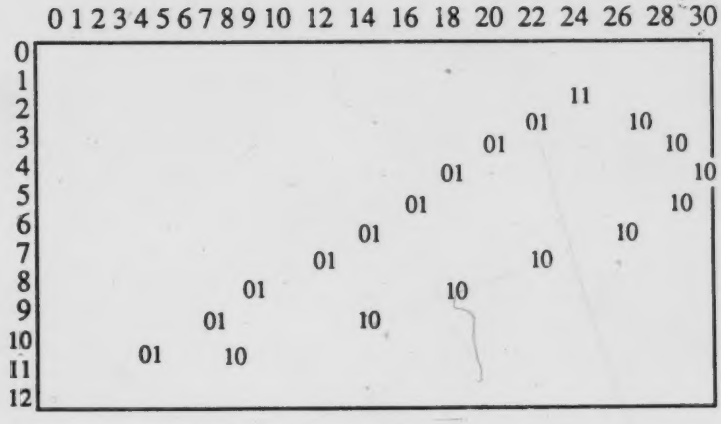
(3) transform the flag matrix to a dot matrix for output and store it in the laser-scan buffer.

Fig. 6.4 illustrates the steps for a regular stroke.

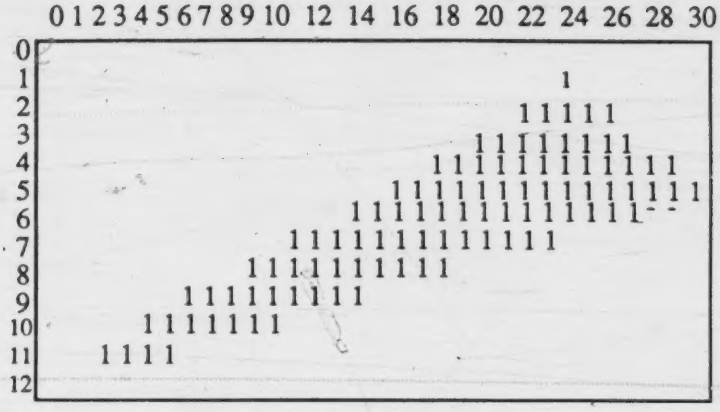


(a) a series of vectors and the contour dot matrix

(* is the starting point of a vector)



(b) the flag dot matrix (two bits per pixel)



(c) the final dot matrix for output

Fig. 6.4 The Contour Matrix, Flag Matrix and Final Dot Matrix for Output [GUO85]

The three steps above are implemented by two high speed microprocessors. In order to achieve high speed and high quality for output, the compression method for Chinese information, recovery algorithm, and the logic design of the microprocessors should be chosen (or considered) very carefully.

The input to the recovery device is the compression representation of Chinese characters which is sorted by strokes (i.e. regular strokes are entered first followed by irregular strokes) and the strokes are received one after another. However, the output from the recovery device is the dot matrices which is sorted by scan lines, i.e. the scan lines (not strokes) are generated one after another. Therefore there exists an "order conflict" between input and output data.

The solution to the problem is to buffer the intermediate data. The data structure of this intermediate data is very important. The flag matrix is chosen as the intermediate data structure so that the transformation from the compressed data to the flag matrix and from the flag matrix to a final dot matrix are simple and fast. The contour matrix is not chosen as the intermediate data structure because the algorithm from the contour matrix to final dot matrix is very complex and difficult to implement.

6.3.3. Enlarging and Shrinking of High Resolution Chinese Characters

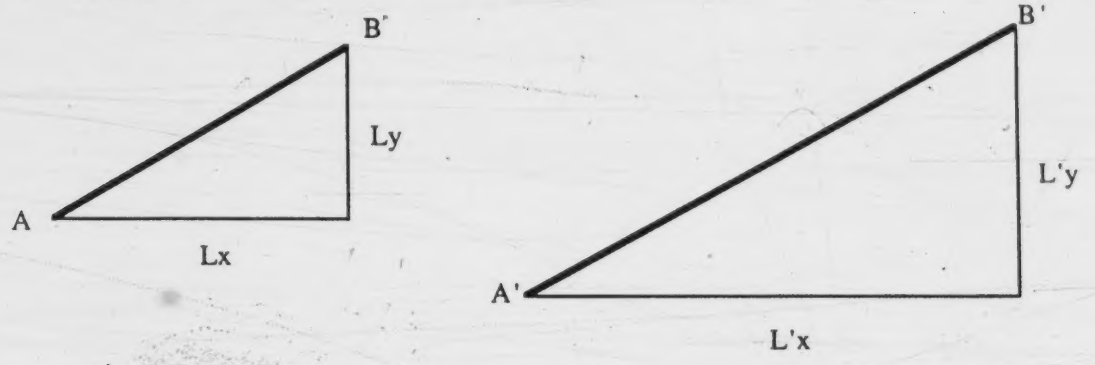
The Hua-Guang system by Beijing University supports characters with 16 sizes (Table VII). The 5-point character is the basis for other characters including Song for books, Song for newspapers, Heading Song, Imitative Song, Bold, and Regular. The font library stores 5-point characters only and all other sizes are generated by enlarging or shrinking the 5-point characters. Besides, four very large sizes of characters are also stored in the font library which are Large

Heading Song, Large Bold, Li and New Wei.

The fundamentals of enlarging and shrinking Chinese characters is simple due to the fact that the strokes are represented by polylines. To enlarge (or shrink) a polyline *r* times, a straightforward way is to enlarge (or shrink) each line segment *r* times. However, in order to prevent the characters from possible distortion during the changes, the following approaches are adopted:

(1) Eliminate accumulative error

For irregular strokes, a method called the node-method is adopted to prevent distortion caused by accumulative error. The formula to obtain the length of a line segment after enlargement is given by the following:



for a line segment AB with

Lx --- the length along X-axis

Ly --- the length along Y-axis

the new line segment A'B' with L'x and L'y can be obtained by

$$L'x = X_{B'} - X_{A'}$$
$$= X_B * r - X_A * r$$
$$L'y = Y_{B'} - Y_{A'}$$
$$= Y_B * r - Y_A * r$$

where $X_{B'}$, $X_{A'}$, $Y_{B'}$, $Y_{A'}$ are X and Y coordinates for A' and B' respectively and X_B, X_A, Y_B, Y_A are X and Y coordinates for A and B respectively. r is the enlargement factor.

(2) Guarantee consistent widths of regular strokes

The node-method may cause inconsistent width for regular strokes. For example, two horizontal lines with the same width may become two new lines with different widths after enlargement. The formula for enlarging regular strokes is:

$$W' = W * r$$

where W' is the new width of a stroke after enlargement,
W is the width of the original stroke,
r is the enlargement factor.

(3) Control the width of horizontal lines for characters with small sizes

Song style characters may be shrunk from 5-point to 6-point or even 7-point characters. In this case, to obtain clear strokes and horizontal lines with consistent width, two bits are used to describe the width of horizontal lines in the following way:

based on a 5-point character (i.e. 96 * 96 dot matrix),

00 --- 1.7 dots (or grid)
01 --- 2 dots (or grid)
10 --- 2.6 dots (or grid)
11 --- 3 dots (or grid).

To obtain high quality characters for all sizes is the central problem to be solved in Chinese character compression. The methods introduced above can hardly be implemented by conventional methods (e.g. medium scale IC). An AM2900 microprocessor is chosen to achieve high speed and low cost because the format and content of its micro-instruction can be determined by the system designer.

During enlargement (or shrinkage) of strokes, multiplication is necessary to obtain new X and Y coordinates. The AM2900 microprocessor is slow for multiplication and even a dedicated multiplier is not fast enough. Therefore, a PROM is used to replace the multiplication operation. The PROM stores the results of multiplication for all X and Y values (i.e. 0 ~127). For each enlarging-factor r, a 128 * 8 PROM is used with its address as the original X,Y coordinate and its contents the results of the multiplication. The enlargement can then be performed in one clock cycle.

6.4. Character Generating by Segments and Complex Layouts

In the previous section, a high quality character generating technique with high compression ratio is introduced mainly from the theoretical point of view. In this section the hardware implementation aspects (i.e. the AM2900 microprocessor) will be discussed which is the key issue for the HuaGuang Chinese publishing systems.

6.4.1. Two Level Storage and Management for High Resolution Characters

At the beginning of the chapter, the features of laser typesetters were introduced. One major problem is the complicated control due to the fact that during scanning the laser typesetter is not allowed to stop in the middle of a page, therefore the controller must guarantee to provide information for the layout dot matrices at any moment. To meet the requirement of laser typesetter a two level storage and management system for character typeface is adopted in the system developed by Beijing University (Fig. 6.6).

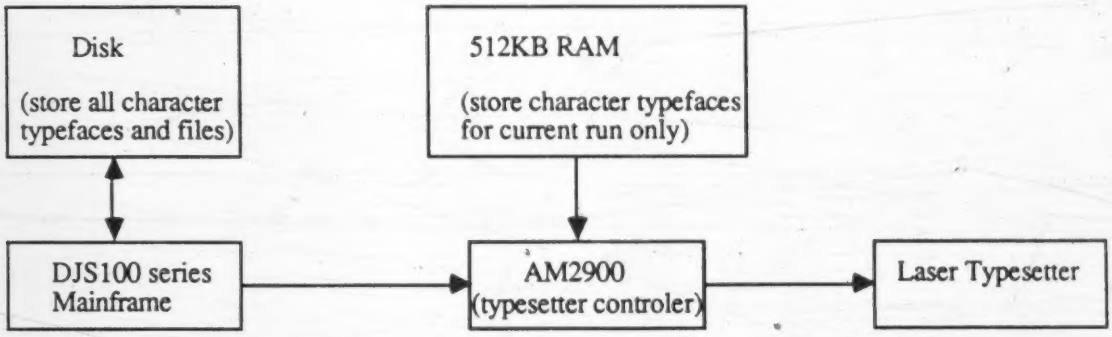


Fig. 6.5 Two-Level-Storage of Chinese Character Typefaces [GUO85]

The purpose of the two-level-storage is not to access disk during the entire typesetting process (e.g. typesetting a whole book or a newspaper) because it takes 10 to 100 ms to fetch a character from the disk. Figure 6.5 indicates that Chinese character typefaces are stored on two levels, namely, the disk that stores all character typefaces and the 512KB RAM that stores only those typefaces needed in current typesetting.

Before output begins, the DJS100 main frame selects all typefaces needed by this typesetting

and stores them in the 512KB RAM of the typesetter so that it can output the whole book or page of a newspaper without accessing the disk. Note that although there are huge number of Chinese typefaces including more than ten styles with 6,000 to 10,000 characters for each, the number of typefaces for one specific book or newspaper is not large (usually less than 4,000). With a high ratio compression technique a character dot matrix needs only 100 bytes. As a result, 512KB is enough to hold all character dot matrices for a specific typesetting job.

6.4.2. Generating Character Dot Matrices by Segments

As mentioned before, the laser typesetter generates output based on scan lines while the high compression technique describes characters based on strokes, so there exists a so called "order conflict". To solve the problem, character dot matrices are generated by segments.

The scan-buffer in the typesetter stores two segments each time with 32 lines for each segment. A character recovery device generates only one segment (32 lines) at a time. For example, to generate one line of N Chinese characters, the order in which the character dot matrices are generated is the following:

generate the first segment of the first character, then generate the first segment of the second character, and so on, until the first segments for all N characters are generated and stored in one part of the scan-buffer; when the contents of the other part of the scan-buffer has been output, then generate the second segment of the first character, then generate the second segment of the second character, and so on, until the second segments for all N characters are generated, and so on.

6.4.3. Complex Layout Generation

The information to describe a publishing layout accurately is called editing-information which is generated by software on a main frame and provided to the typesetter. The typesetter then generates the layout dot matrices based on the information.

The editing-information describes the layout for a line which consists of the following 6 indicators:

- Line start flag: indicates the number of segments in the line;
- Character style flag: indicates the style and the sizes of X, Y directions. If X and Y are equal, a square character is generated, otherwise, a long or flat character is generatedaccording to the aspect ratio specified by X and Y;
- Internal code flag: the address of a character in the 512KB RAM can be found by the internal code flag;
- X-location flag: indicates the X-location on the layout for a character dot matrix;

- Y-location flag: indicates the Y-location on the layout for a character dot matrix and if the characters following have the same Y-location the flag is shared by the character string;
- Line end flag: indicates the end of the line, or a page, or typesetting.

Note that the Y-location flag indicates the difference between the current line and the first row of the character dot matrix. If this value is negative, then the character is located below this line and if the value is positive, the character is above the line. This flag is useful for complex layouts such as vertical headings, mathematical formulas, tables and graphs. By using the above 6 flags any complex layout can be accurately described.

6.4.4. Typeset Controller

The typeset controller consists of two AM2900 microprocessors. The microprocessor G has 2K micro-instructions with a length of 56 bits each and the microprocessor F has 512 micro-instructions with a length of 48 bits each. Both of them have a 16 bit ALU (4 AM2901A for each) and a microprogram controller (1 AM2910 Fig. 6.7).

The function of the microprocessor G is to:

(1) fetch a line of editing information from the main memory of the DJS100 main frame and store it in local memory;

(2) process the various flags of the line editing information one by one;

(3) implement the first two steps of character recovery, i.e. convert the compressed representation of regular or irregular strokes to the standard vector string and transform the vector string to the flag dot matrix.

The function of the microprocessor F is to implement the last step of character recovery, i.e. convert the flag dot matrix to the final dot matrix for output, and to control the laser scan (Fig. 6.6). The details of figure 6.7 will not be described here and only the key points are given below:

(1) microprocessor F sends the final dot matrices to the appropriate position of the scan-buffer SS according to the flag X-location provided by microprocessor G;

(2) the F also controls laser scanning by fetching 4*4 bits of information from SS and sending it to the laser typesetter (each time the laser typesetter finishes scanning the 4*4 dots it sends a request signal to F);

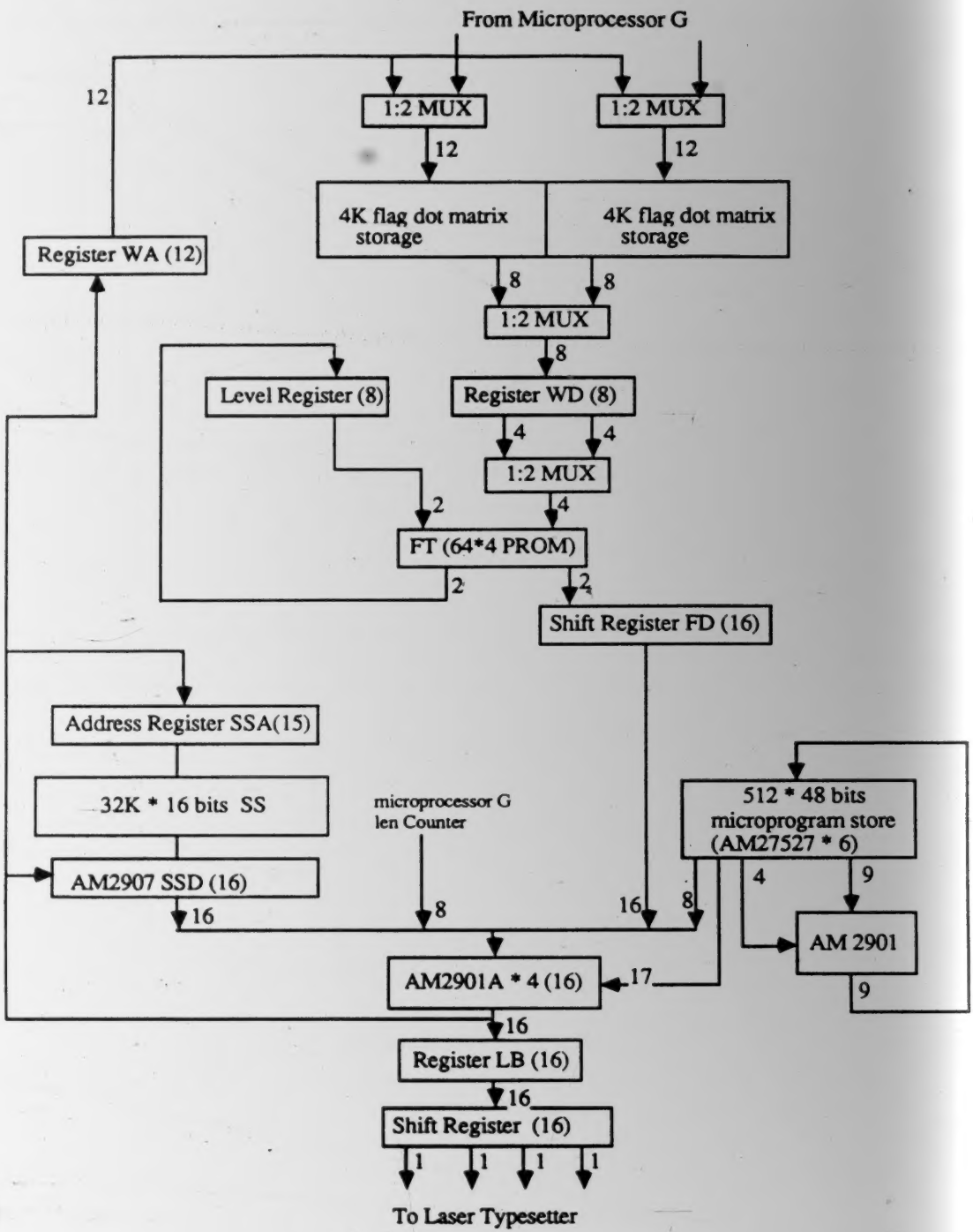


Fig. 6.6 Microprocessor F [GUO85]

(3) the flag dot matrix is stored in the 4K memory line by line, while SS stores the final dot matrix 4-lines by 4-lines, i.e. 16 bits of each unit belongs to 4 scan lines. The flag dot matrix should not be converted line by line and the Level Register ensures the conversion is performed properly.

6.5. The Whole Hua-Guang Publishing System

The brief description of the Hua-Guang publishing system is given in this section as a typical Chinese Electronic publishing system.

6.5.1. Hardware

Figure 6.8 shows the hardware of Hua-Guang system. The main frame is a DJS153 minicomputer (made in China) or a NOVA3 (or NOVA4) minicomputer with each main frame connected to 6 to 8 Chinese terminals. Each terminal has a Chinese dot matrix printer. A hard disk with more than 20MB capacity stores high quality Chinese characters, compressed representation of each character styles, system software, typesetting documents and intermediate results. The typesetter controller is connected to the main frame as an auxiliary device which is responsible for converting the compressed information to final dot matrices, controlling the laser typesetter to print dot matrices on the proper position of a film, and controlling the full size output device to print a full size newspaper page.

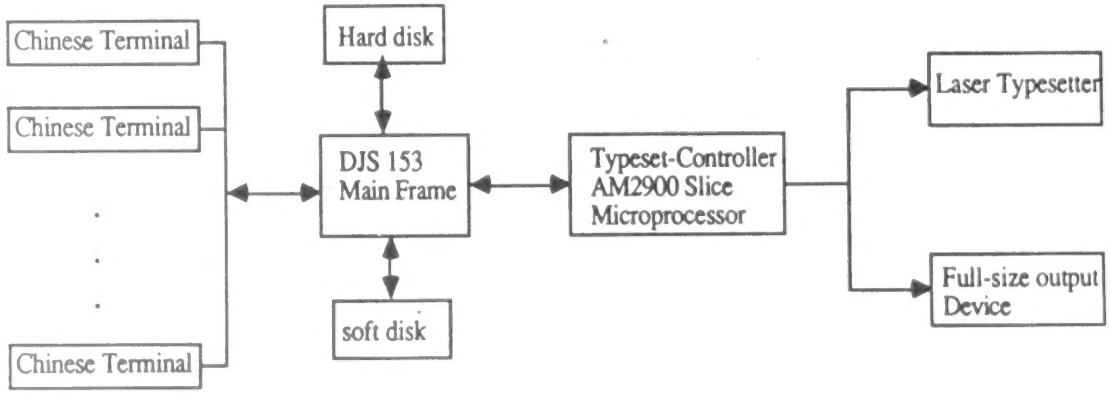


Fig. 6.7 The Hardware Configuration of Hua-Guang System [GUO85]

6.5.2. Working Procedure of the Editing-Publishing System

First, the user inputs the document with the publishing format requirements (called a galley proof in the publishing industry) from a Chinese terminal and store it on a floppy disk. After

editing the document and the format requirements on the terminal the user issues an input command and transfers the galley proof from the floppy disk to the main frames hard disk.

After all galley proof is sent to the main frame, the user issues a typesetting command from the terminal and the main system generates the output for the typesetter according to the format requirement such as character style, size, position, etc. The main system also gives "typesetting finishing" information on the terminal, e.g. the number of pages for a book or display part of the layout. The main system can also give error messages if the user's format requirement is not correct and display the document on the terminal for the user to correct the mistakes until a satisfactory layout is obtained. The whole process takes only a few minutes.

The user can issue "full size output" command from the terminal to ask the output device to generate a full size layout of a newspaper (exactly the same size as a page of the final newspaper). After the user examins the whole page and is satisfied with it, the typesetting command for the typesetter is issued to output the layout onto film for publishing.

The typesetting language for Chinese publishing is called CL Typesetting Language (or CL Language) which is used mainly for typesetting books, magazines and newspapers. It consists of typesetting annotation and characters (symbols).

(1) Typesetting annotation

Typesetting annotation is similar to statements of any programming language that describes various typesetting requirements. The format of the typesetting annotation is:

[(annotation name) [(parameter)]]

For example, 中[KG1.3]加友谊 ("China/Canada friendship" in English) means that the three characters following 中 will have an interval of one character between them, i.e.

中 加 友 谊

where KG is the annotation name which means "space" and 1.3 is the parameter which means to leave a space of one character for the following three characters.

(2) Characters (symbols)

The characters (or symbols) in the CL Language are just like variables or data in any programming language. The only difference is that the CL Language provides Chinese characters.

(3) CL typesetting compiler program

The CL compiler program includes three scans of the source: the first scan processes the galley proof on terminals, namely, check syntax errors, transfer the galley proof into an intermediate language and transfer the annotation into an intermediate form; the second scan compiles the output from the first scan, i.e. determines character style, size and its position on layout, also realizes automatic line-breaks, page-breaks, etc.; and the third scan processes the result from the second scan and generates the form suitable for a laser typesetter.

(4) Two examples of CL language application

Example 1: To generate the following characters with different sizes ("Chinese Editting/Typesetting" in English).

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

汉字编辑排版

where the character sizes are different for different rows, the row interval is 1/2 and the character interval is 1/3 of the character size on that row respectively. All rows are centered. The typeset-annotation is written as:

[HJ 2:1] [HT 0 DH] [JZ (*3] 汉字编辑排版 [HT 1 DH] 汉字编辑排版 [HT 2 DH] 汉字编辑排版 [HT 3 DH] 汉字编辑排版 [HT 4 DH] 汉字编辑排版 [HT 5 DH] 汉字编辑排版 [JZ)]

where [HJ 2:1] indicates the row interval, [JZ (*3] [JZ)] indicates that the characters are centered (*3 is the interval between characters), and the numbers from 1 to 5 are the characters sizes.

Example 2: Newspaper layout design

There are two main annotations for newspaper layouts: the section description and the headline description.

An example of a section description is:

[FQ2B: C1304 (0101Q41*10+25*10.3 [Z4*3L]) C1305 (X42712Q16*10.3) C1306 (0145Q41 * 21[Y23 * 4]) C1307 (4401Q33 * 21.2 - ZX10*10+33*21 [Z4*2L; YX 3 *1L])].

This is the second page of an octavo newspaper consisting four articles. Their file names are C1304, C1305, C1306 and C1307. Among them C1304 is typeset starting from the first character of the first row and it consists four columns with the first column containing 41 rows (with 10 characters for each row). Both second and third columns contain 25 rows. The headline is at the upper left corner of the article which takes the area of four rows, etc. The other three articles are described in the same way. Figure 6.9 shows the resulting page layout.

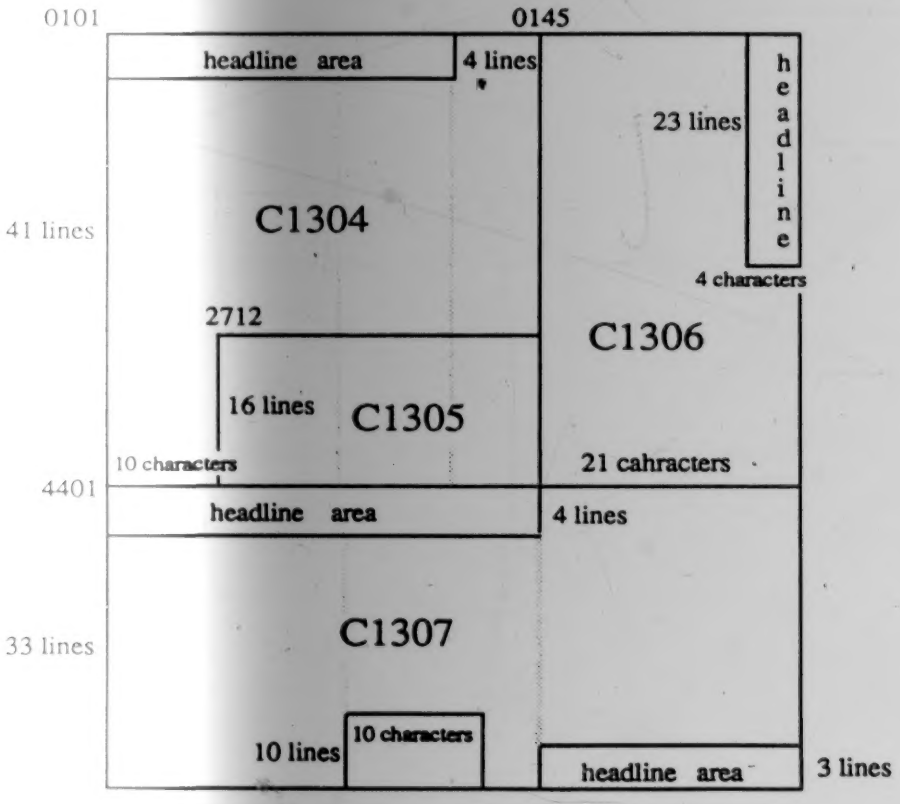


Fig. 6.8 Layout of the Example Annotation [GUO85]

6.6. Other Systems

A brief survey of other electronic publishing systems is given in this section which includes typesetting systems (for books and newspapers) and desk top publishing systems.

6.6.1. The Ke-Yin and 4S System

The Ke-Yin and 4S system was developed by the China Printing Technology Institute and Stone Corporation. As described in Section 6.2, a run length code has the disadvantage of a low compression ratio for representing Chinese characters. The Ke-Yin and 4S system aims at improving the run length coding technique in that an enhanced run length coding is used with a data compression ratio of 8 times higher than that of a conventional one. The total data compression ratio is 200. The MONOTYPE MARK II laser typesetting machine used in this system has a resolution of 1,000 lines/inch. For typesetting Chinese characters, a 200-300 MB high speed disk should be used with the typesetting machine. 1380 * 1380 dot matrices are used to represent Chinese characters, so the printing quality is high. This system is capable of typesetting books and octavo newspapers without cuting and pasteing. There are already more than twenty installations of this system in China [YU88].

6.6.2. The UDN (United Daily News) System

The UDN system was developed by the United Daily News of Taiwan. In this system, an Autologic APS-41 scanning system is used to create Chinese fonts. Chinese characters are first photographed on film and then fed into the system for a fine scan with a resolution of 1,400 lines/inch. The scanned character is edited, compressed and stored onto a 300 MB hard disk. The Autologic APS-5 CRT phototypesetting system is used to print formatted fine quality masters for newspaper page paste-up and plate making. The output speed is 100 Chinese characters per second at a resolution of 1,000 or more lines per inch.

This is a very large system. It produces not only daily UDN's six world-wide Chinese newspapers but also volumes of periodicals and books. The publishing capacity of this system is two million Chinese words each day, so the reliability of the system is very important. To provide high reliability and contingency operation, recovery procedures are built into the system software. During newspaper production, two C/350 computers are run in tandem and a third is on standby [YU88].

6.6.3. The NewStar Computer Chinese Typesetting System (NS System)

Unlike the UDN system which is a very large and advanced typesetting system, the NS system

is a small and popular typesetting system which was developed by the NS Electronic Typesetting Corporation (Hong Kong), the Shang Hai Printing Technology Institute (China) and some other organizations.

The hardware of the NS System consists of an IBM-PC computer, the NS-400 and ZPB-256 typesetter and a laser printer. The NS System supports 4 basic styles of Chinese characters with 24 sizes and it can also support more than 10 extra styles including the Li style. The typesetter can output 256 * 265 mm^2 and 265 * 400 mm^2 films with a resolution of 750 lines/inch and 1000 lines/inch. A compression technique makes it possible to store all high quality character fonts in the microcomputer so that the NS system obtains a very good cost/performance ratio (the basic system costs only about $ 25,000 Canadian dollars). The system has been installed and used in the Canadian Chinese Times Cor. (加华日报) and Chinese Computer Typesetting Center Inc. in Toronto (Canada), QingHua University, KunMing Industrial Institute, YunNan University, Foreign Language Publishing House, Economic Information Newspaper, ShangHai Printing Technique Development Corporation, etc.

6.6.4. Desk Top Electronic Publishing Systems

The Desk Top Publishing (DTP) system is used to generate memos, reports, posters, internal newsletters, technical documents, and advertisements by many organizations. The primary consideration is the cost of the DTP system and the cost of each page printed. Moderate printing quality is sufficient for most applications. The printing resolution is normally at 300 lines/inch. Some typical DTP systems are briefly introduced bellow:

(1) The Bi-Sheng System

The Bi-Sheng System was developed by the Changsha QianJin (Advanced) Computer Institute. In this system, a special Chinese character generator called the Bi-Sheng Chinese Card is used. This card can generate 13 typefaces, each contains 7,000 Chinese characters. The character sizes can be varied continuously from 16 * 16 dots to 384 * 384 dots. The memory requirement for storing the compressed data is only 350 KB. There is no CPU on the Bi-Sheng card. The CPU of the PC is used to generate the Chinese characters. The generation speed is 80 Chinese characters per second for characters of 48 * 48 dots. Based on the Bi-Sheng Chinese card, the Institute has also developed their own Chinese character operating system, QJCDOS, the editing and composing software [YU88].

(2) The Advanced Document Typesetting/Printing System (ADTPS)

Since the GreatWall 0520 (a Chinese IBM-PC compatible microcomputer) series is very popular in China, many software application (including DTP) have been developed for it and ADTPS is such a DTP system developed by the Chinese People's Liberation Army for the purpose

of office automation. A Continuous Smoothing technique is adopted to guarantee printing quality for different sizes of characters. The system supports four styles with four sizes for each and users can change style and sizes of characters at any part of a document. The system includes Table Generating software which can describe a table format by simple commands in general editing mode (e.g. WordStar or EDLIN in the IBM-PC). It can print smooth lines with any angles, and change the style and sizes of characters within an entry of a table. The ADTPS have been installed in some offices of the Chinese army and navy.

(3) The HG-2000C system

The HG-2000C system was developed by the WeiFang Computer Corporation. No special hardware is required for this system. Dot matrices of Chinese characters for different typefaces and sizes are stored on a hard disk. The company has developed its own editing and composing software.

6.7. Conclusion

In this chapter, the Hua-Guang system developed by Beijing University was introduced in depth in order to show the main features of a Chinese electronic publishing system. The main technical difficulties faced by Hua-Guang system (also by all Chinese publishing systems) are: huge typeface storage, fast access, high quality character output, and complicated layout requirements (which needs powerful software).

With the development of VLSI circuits, one card may be enough to contain a large Chinese font library, character recovery device, and typeset-controller, so that the typesetting system may become an ordinary device in people's offices. No-film laser typesetting is called the 5th generation typesetter which eliminates all work involved in film processing, improving efficiency and reducing cost. Connecting a typesetting system to micro-film devices for publishing archives and connecting it with an information retrieval system for publishing retrieval are also important research areas.

In Chapter 8, a concept of integrated font generating scheme is discussed where a high quality character generation method is introduced as a compensation to a main font library.

Chapter 7
AUTOMATIC PROCESSING OF THE CHINESE LANGUAGE

Automatic Processing of the Chinese Language (APCL) is a higher stage of the development of Chinese Computing due to the following observation: at the primary stage of the development of Chinese Computing, Chinese input/output and Chinese/English compatibility are bottleneck for obvious reasons. After these bottleneck problems are solved, it is very natural for people to ask for more advanced application, the APCL. This is because: (1) APCL helps to design an advanced Chinese computer system or improve the available Chinese systems (e.g. finding an input method acceptable worldwide) and a computer-aided language processing technique is necessary for analyzing the features of Chinese and obtaining important statistical results, (2) APCL systems are indispensable for establishing Chinese database, Chinese knowledge base and Chinese information base which are the main goal of Chinese Computing. For years, computers have been used for applications related to natural language processing besides the scientific and engineering applications, which is the combination of word processing, pattern recognition, and artificial intelligence. APCL is a very active research area in this field.

A language processing system consists of four parts: a language processing software system, a language knowledge base, a language information base, and a language data base (Fig. 7.1). The language knowledge base stores rules and modules for a natural language. The language information base stores all kinds of "raw" language materials such as documents, novels, poems, etc. The Language Processing Software System processes the "raw" language materials and outputs well organized data stored in the language data base for users to retrieve.

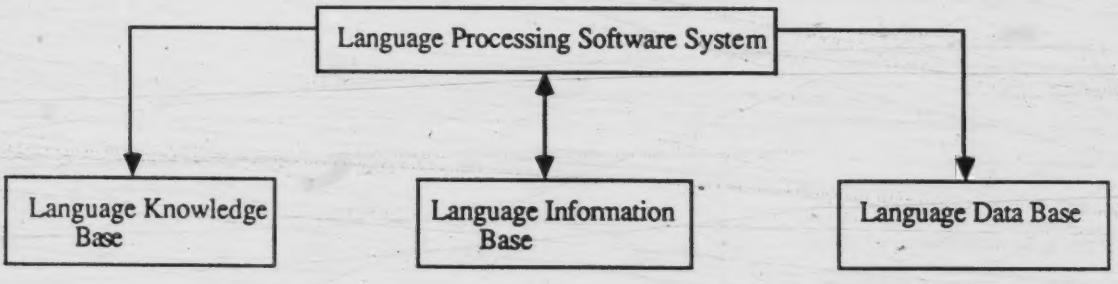


Fig. 7.1 The Structure of Language Automatic Processing System [GUO85]

The central part of an automatic processing language system is the language processing software system (or program library) which is capable of:

- written and oral language processing;
- automatic conversion between different natural languages (i.e. machine translation);
- processing a language according to chronicles, e.g. modern and ancient Chinese processing;
- processing all aspects of a natural language, such as phonetics, vocabulary, grammar, etc. (grammar processing is the most difficult among them).

Machine translation is a very important part of APCL because (1) China is a multi-national country, a APCL system should be capable of processing different languages and performing translation between them (e.g. Tibeten and Hanzi information) and (2) Many relations are developed between China and most countries in the world and most developed countries do not use Chinese. Therefore, the state of art of machine translation is introduced here.

7.1 Difficulties in Machine Translation

Owing to its potential in commercial, cultural and educational values, machine translation has attracted active researchers even in the early sixties. In spite of many years of research by various groups, so far there is no single machine translation system that can claim to achieve its original objective, namely, the capability of automatic translation (without human help) of text (general knowledge domain) from one language to another. The failure to achieve such a goal is partly due to the limitation of computer hardware (i.e. speed and memory size) but mostly due to the complexity of language. It is now realized that a well developed language such as English or Chinese is so complex that its structure cannot be effectively described by a finite set of syntactic rules. Also, it is now realized that the "meaning" hidden in a given sentence comes partly from the (dictionary) meaning (which is not unique) of each word, partly comes from the context of other sentences and partly comes from the domain knowledge and cultural background of the reader. Obviously, computers lack cultural background and domain-knowledge. Hence, it is clear that, with present computer technology, it is impossible to develop a general purpose, totally independent machine translation system between well developed languages such as English and Chinese.

7.2. Available Computer Translation Systems

Owing to a deeper understanding of the nature and difficulty of machine translation,

researchers of today are much more realistic. The objective of machine translation systems is now redefined as a computer program (or a family of programs) to assist human translators to be more productive. This implies that machine translation systems should be designed on a much narrower domain. Also, they may require human translators to do a final touch up or even preprocessing of the text to be translated.

Under the modified objective, there are in fact many machine translation projects actively engaged by researchers all over the world. This is clearly seen from the recent Machine Translation Summit (Sep. 17-19, 1987) in Japan. There are even commercially available machine translation systems developed by various computer companies which are on exhibition at some Chinese processing international conferences [YU88].

China has produced its first commercial system called "译星一号" (Translation Star-I) which is a Chinese/English machine translation system developed by the China Software Technology Corporation. The system includes a word base with more than 100,000 English words and special purpose dictionaries for economics, communications, computers, etc. The correctness of translation is 80 % (i.e. English text directly input into the system without pre-processing by human). The system can be used on many popular computers such as the GreatWall 0520, IBM-PC, Wang, and 68000. The speed of translation increases with the computers capacity.

China and Japan have also a joint project of developing a machine translation system between Chinese and Japanese. China and France (University of Genoble) are planning to jointly develop a Chinese/French machine translation system. At the Institute of Systems Science in Singapore, there is also a small project of developing English/ Chinese machine translation using bilingual PROLOG and bilingual LISP on PCs and workstations.

7.3. Conclusion

Automatic Processing of the Chinese Language (APCL) is a higher stage of the development of Chinese Computing. The structure of APCL systems is briefly introduced in this chapter. The central part of the APCL system, the machine translation, is introduced in terms of its difficulty together with some available machine translation systems.

Chapter 8
SUMMARY OF RESULTS AND RECOMMENDATIONS

So far, the most important areas of Chinese computing have been studied in Chapters 2 to 7. Based on these studies, three recommendations are given: the Character Decomposition Pinyin (CDPinyin) which is an improvement on TianMa CWP, the Character Channel which is a systematic way to solve the Chinese/English compatibility problem, and High Quality Chinese Character Generating for Casual Users, an experiment to investigate the feasibility of generating high quality character using popular laser printing facilities.

8.1. Character Decomposition Pinyin (CDPinyin)

In chapter 5, the advantages and disadvantages of the Pinyin input method are discussed and one of its major disadvantages is that a user cannot use Pinyin if the character to be input is not known. In this case the user has to either look up in a dictionary to find the Pinyin (pronunciation) of the character or use other shape-based input methods. Obviously, both approaches can hardly be accepted by ordinary people, because looking up in a dictionary by components of a character is very time-consuming and learning any shape-based input method will take a considerable amount of time. CDPinyin is introduced to tackle the problem. The idea behind CDPinyin is to help users input unknown characters by a phonetic method so that they need neither to learn a shape-based method, nor to look up in a dictionary.

8.1.1. What is CDPinyin?

CDPinyin is introduced as an improvement on TianMa CWP in order to solve the problem mentioned above. The method is based on the following observation:

Although ordinary Chinese people know only three to four thousands characters, most characters are new (i.e. they cannot pronounce the character) but they consist of simpler characters or basic components which can be pronounced. In fact, when people try to express a new complex character (e.g. zhang 樟) they often express it as the composition of several simpler components (e.g. 樟 = 木+ 立+ 早, "Zhang is Mu plus Li plus Zao" in English). CDPinyin (Character Decomposition Pinyin) works exactly the same way, that is, for each

Table XVI. Some Examples of CDPinyin.

	Position	Pinyin	Decomposition
	5745	sang	磉 = 石+桑 = shi + sang
	5746	piao	瞟 = 目+票 = mu + piao
	5747	yan	罨 = 四+奄 = si + yan
	5748	lan	罱 = 四+南 = si + nan
*****	5767	tang	耩 = 耒+尚 = lei (geng) + shang
* *	5778	you	蝣 = 虫+游-水 = chong + you - shui
*	5779	fou	缶
	5780	shi	舐 = 舌+氏 = she + shi
	5781	jian	笕 = 竹+见 = zhu + jian
****	5092	zha	吒 = 口+丿+七 = kou + 3 + qi = 口+托-手 = kou + tuo - shou
***	5300	shu	倏 = 人+丨+文+犬 = ren + 2 + wen + quan
	4863	hu	壶 = 士+冖+业 = shi + tubao + ye
*****	5385	tang	醣 = 酉+唐 = you + tang = jiu + tang
	5801	luo	雒 = 各+淮-水 = ge + huai - shui
	5802	wu	鋈 = 沃+金 = wo + jin
	5804	yang	鞅 = 革+央 = ge + yang

Notation and Explanation for Table.XVI:

* indicates that the character can not be decomposited, e.g. 缶

** indicates that the character-decomposition involves "-" which means to add a character first and then remove the extra radical, e.g. 蝣

*** indicates that the character-decomposition involves one of the five basic strokes, (丨 丿 乀 乛) that are represented by 1 to 5 respectively. The five basic strokes are taken form Mr.Wang Yongmin's Five Stroke Method, e.g. 倏 and 吒.

**** indicates that the character may be decomposited by more than one ways depending on user's habit, e.g. 吒.

***** indicates that the radicals of a character may be pronounced in different ways depending on the user, e.g. 鹅 and 醣 .

character the input method provides two ways to pronounce it : to pronounce the character as a whole or to pronounce it as the composition of several components. For example, "樟" may be pronounced as "Zhang" or "Mu-Li-Zao" and a user can get the character "樟" in either way.
There are some special cases that CDPinyin has to be capable of dealing with:
Case 1: the character cannot be decomposed easily.
Case 2: the character-decomposition requires both "addition" and "subtraction" e.g. 龈 = 齿 + 银 - 金 ("Yin is Chi plus Yin minus Jin" in English).
Case 3: the character-decomposition involves five basic strokes (i.e.一 丨 丿 乀 乛).
Case 4: the character-decomposition involves some components which may be pronounced in different ways.

8.1.2. Applying CDPinyin on a Subset of the Hanzi Frequency of Usage for News Information (新闻字频表).

The Hanzi Frequency of Usage for News Information [NEWS87] is a Chinese character

table established by the China News Agency in which 6000 characters are sorted according to their frequency of use in newspapers in China. The last 1000 characters in the Hanzi Frequency of Usage for News Information [NEWS87] are used to test the validity of CDPinyin method and the reason for choosing these characters is that:

(1) These characters are seldom used and probably are new characters to most users.

(2) A set of 1000 characters is big enough to draw some meaningful conclusions for the CDPinyin method. Figure 8.1 shows the results of applying CDPinyin to some characters in the character table. Special cases are marked by "*" and this implies that some extent of intelligence is required for CDPinyn to handle these cases.

The results are the folowing:

Among 1000 characters (from 4810 to 5809 in the Hanzi Frequency of Usage for News Information) 37 characters can hardly be decomposed, 32 characters need "-" (subtraction), 4 characters involve five basic strokes, and 22 characters involve multi-name radicals (depending on user habits).

(1) The following characters can hardly be decomposed and users have to look up in a dictionary for their Pinyin (pronunciation):

夔 (kui 4817)	睿 (rui 4849)	肄 (yi 4886)	盥 (guan 4935)	臾 (yu 4947)
豢 (huan 4964)	戮 (lu 4966)	誊 (teng 4968)	捋 (lu 4987)	辎 (zi 5017)
闶 (hong 5106)	觞 (shang 5316)	餮 (tie 5141)	片 (pan 5211)	虢 (guo 5224)
毋 (bi 5227)	蠹 (du 5246)	丐 (gai 5295)	刈 (yi 5299)	禀 (bin 5301)
舛 (chuan 5392)	硇 (nao 5361)	匜 (yi 5394)	叵 (po 5423)	奘 (zang 5446)
桀 (jie 5493)	塍 (cheng 5504)	篦 (bi 5573)	彧 (yu 5573)	黍 (shu 5589)
亓 (qi 5602)	弁 (bian 5621)	畚 (ben 5622)	葑 (hong 5639)	缶 (fo 57799)
笸 (po 5785)	羿 (yi 5791)			

Ratio: 37 / 1000 = 3.7 %

(2) The following character-decomposition involves 5 basic strokes (一 丨 丿 八 乛) which are represented by 1 to 5 respectively.

虬 (虫 + 5 qiu 5042)　　吒 (口 + 3 + 七 zha 5092)
虱 (3 + 虫 + 5 shi 4812)　　倏 (人 + 2 + 文 + 犬 shu 5300)

Ratio: 4 / 1000 = 0.4 %

(3) The following character-decomposition involves "-" (subtraction), i.e. it needs to remove an extra part from an added component:

媪 (女+温-水 ao 4835)　　槿 (木+谨-言 jin 4840)　　诿 (言+橘-木 jue 5075)
畝 (插-手+欠 sha 4845)　　刎 (忽-心+刀 wen 4972)　　饬 (食+伤-人 chi 5206)
荒 (经-丝+流-水 qiu 5079)　　狻 (犬+俊-人 suan 5203)　　羧 (羊+峻-山 suo 5250)
逢 (走+降-耳 pang 5216)　　瑗 (王+援-手 yuan 5217)　　皴 (俊-人+皮 cun 5370)
龈 (齿+银-金 yin 5255)　　爰 (援-手 yuan 5356)　　癸 (葵-草 kui 5411)
羯 (羊+渴-水 jie 5379)　　粼 (磷-石 lin 5405)　　婕 (捷-手+女 nu 5484)
匍 (葡-草 pu 5430)　　怿 (心+译-言 yi 5472)　　畯 (田+俊-人 jun 5583)
觖 (角+块-土 jue 5552)　　嫽 (女+谬-言 lao 5569)　　谖 (言+援-手 xuan 5612)
骢 (马+聪-耳 cong 5697)　　谄 (言+掐-手 chan 5610)　　祓 (示+拔-手 fu 5739)
蝣 (虫+游-水 you 5778)　　殄 (歹+珍-王 tian 5719)　　雒 (各+推-手 luo 5810)
赦 (赤+报-手 she 5793)

Ratio: 31 / 1000 = 3.1 %

(4) Some radicls may be named in different ways depending on user habits. Table XVII shows some examples.

Table XVII. Some Radicals (or Characters) with Different Names

character (or radical)	Pinyin	name commonly guessed by Chinese people
豕	shi	zhu 猪
酉	you	jiu 酒
豸	zhi	bao 豹 or chai 豺
耒	lai	geng 耕

Among 188 radicals in New-China-Dictionary (XinHuaZiTian), 164 of them can be pronounced by users without confusion, but some may be called in a simplified way and the system should allow users to pronounce the radicals in whichever way they want. Some examples are:

Table XVIII. The Radicals with Simplified Names

radicals	name	possible simplified pronuciation
冫	两点水 liangdianshui	ershui
冖	秃宝盖 tubaogai	tubao
氵	三点水 sangianshui	shui
礻	示补 shibu	shi
衤	衣补 yibu	yi

The following are taken from the set of 1000 characters of the Hanzi Frequency of Usage for News Information (新闻字频表):

壶 (士+冖+业 shi+tubao+ye 4863)

霾 (雨+豸+里 yu+zhi+li 4950) or bao

貘 (豸+莫 zhi+mo 5135) or bao/chai

滹 (水+虎+乎 shui+hu+hu 5335)

醣 (酉+唐 you+tang 5385) or jiu

酢 (酉+乍 you+zha 5542) or jiu

醺 (酉+熏 you+xun 5544) or jiu

祓 (示+友+丶 shi+you+4 5739)

耦 (耒+尚 lei+shang 5767) or geng

酎 (酉+寸 you+cun 5795) or jiu

酞 (酉+太 you+tai 4883) or jiu + tai

耱 (耒+磨 lei +mo 5040) or geng

耜 (耒+吕 lei+lu 5239) or geng

醌 (酉+昆 you +kun 5384) or jiu

燹 (豕+豕+火 shi+shi+huo 5509) or zhu+zhu

酷 (酉+立+口 you+li+kou 5543) or jiu

豺 (豸+才 zhi+cai 5592) or chai

褰 (衣+襄 yi+xiang 5740)

酊 (酉+丁 you+ding 5794) or jiu

酩 (酉+名 you+ming 5796) or jiu

Ratio: 20 / 1000 = 2 %

Total 3.7 % + 3.1 % + 0.4 % + 2 % = 9.2 %

The experiment shows that among 1000 low-frequency-characters (i.e. seldom-used

characters) only 3.7 % cannot be decomposed, 5.5 % need to be treated in special way (i.e. case 1 to 4) and 90.8 % can be decomposed properly. That is, more than 90 % low-frequency-characters can be decomposed by ordinary people, which shows the validity of CDPinyin.

8.1.3. The Experimental Results of CDPinyin

To justify the validity of CDPinyin, two examples (the News and the Poem used in chapter 5) are tested in the following way:

First, some relatively complex characters are chosen as new characters unknown to a user (indicated by bold characters). Then the examples are typed into TianMa CWP. For each example two results are obtained: one is by looking up dictionary for the new characters and the other is by CDPinyin.

Example 1: News (新闻稿)

The following is the text of Pinyin with CDPinyin for 14 chosen characters:

yong ji **zhu+mu+sheng-3** ji chuli de xin **zi+xin**, baokuo shuju, wenzi, tuxing, yuyin deng, qizhong zhuyao de shi wenzi xinxi. woguo shi ge duomin **fang+3+1+shi** de guojia, da duoshu renmin **shi+yu** hanzi, yinci woguo shiyong de zhuyao shi hanzi. xin zhongguo chengli yihou, chengren quanguo ge minzu yi **shuangren+jin-shui** pingdeng, fandui minzu sha wen zhuyi he **quan+jia er+yi** minzu zhuyi, yinci you biyao qufen "hanzi" he "zhong wen de tifa, buneng ba tamen **shui+kun** tong qilai. zhongwen yin baokuo shaoshu minzu de wenzi.

woguo yao tuiguang jisuanji yingyong, bixu shi geminzu de wenzi dou neng zai jishuanji shang shixian shuru he shuchu. zhexie minzu de wenzi zhong, zhiyou hanzi shi yi zhong xiangxing wenzi, qita de dou shi pinuyin wenzi. pinyin wenzi de zi **zhu+cun** zui duo bu chaoguo shu shi zhong, er xiandai woguo shiyong de hanzi **ren+you** changyong bufen jiu you liu, qi qian zi, erqie zi **xing+tu 3+1+ri+wen** za, zhe jiu shi jisuanji shixian hanzi xinxi chuli cunzai **che+jiao** dou kunnan. reguo **jiao+dao+niu** jue le hanzi xinxi chuli de nanti women qita minzu de wenzi chuli jiu bijiao rongyi jiejue.

Result:

Ratio: 14 characters are chosen as unknown characters,

(new_character) / (total_number_of_characters)

= 14 / 291

= 5 %

Time cost for TianMa:	12 minutes
Time cost for CDPinyin:	13 minutes
Time cost for looking up dictionary:	20 minutes.

Analysis:

Originally, it takes 12 minutes to input the News into TianMa CWP (i.e. assuming no unknown characters exist). For the text with 5 % unknown characters, CDPinyin increases the time cost by 8 %
i.e. (13-12)/12 = 8%, while looking up in a dictionary increases the time cost by 67 %, i.e. (20-12)/12 = 67%.

Example 2: Poem (诗歌)

The following is the Pinyin text of the poem with CDPinyin for 35 chosen characters:

jiang jin jio libai
yi+kou bu jian huanghe zhi shui tian shang lai, **da+hui** liu dao hai bu **3+1+ri+wen** hui.
jun bu jian gao tang ming jing **jei+xin** baifa, **chao-shui** ru qing si **mo+ri** cheng **yu+shan**.
rensheng deyi **3+3+3+ye** jin huan, mo shi jin **mu+zun** kong dui yue.
tiansheng wo cai bi youyong, qianjin san jin huan fu lai.
heng+4+4+4+4 yang **bao+xin** niu qie wei le, hui xu yi **shi+qian** san bai bei.
shang+jin fu zi, dan qiu sheng, jiang jin jiu, bei mo ting.
yi jun ge yi qu, qing jun wei wo **ren+qing** er ting.
zhong gu **shi+yi+yi+gong** yu bu zu gui, dan yuan chang **you+zu** bu fu **you+xing**.
gu lai **you+tu 2+2+you+bei bi+bai bao+shu bao+mo, xin+huai-shui** you yin zhe liu qi ming. chen wang **cao+1+ri** shi **bao+ri+nu** ping le, dou jiu shi qian **ci+xin** huan **yan+nue**.
zhuren he wei yan shao qian, **shuangren+jing-si** xu **shui+gu** qu dui jun **you+shao**.
wu hua ma, qian jin **jiu+yi**, hu re jiang chu huan mei jiu,
yu **ni-ren** tong **jin-xiao** wan gu **qiu+xin**.

Result:

Ratio: 35 characters are chosen as unknown characters.

(new_character) / (total_number_of_characters)
= 35 / 176
= 20 %

Time cost for TianMa:	7 minutes
Time cost for CDPinyin:	11 minutes
Time cost for looking up dictionary:	33 minutes.

Analysis:

Originally, it takes 7 minutes to input the Poem into TianMa CWP (i.e. assuming no unknown characters exist). For the text with 20 % unknown characters, CDPinyin increases the time cost by 57%, i.e. (11-7)/7 = 57 %, while looking up dictionary increases the time cost by 370 %, i.e. (33-7)/7= 370 %.

The results show the obvious advantage of CDPinyin over consulting dictionary. Since over 90 % of new characters (unknown to most users) can be decomposed, CDPinyin allows users to input Chinese text continually without stopping to consult dictionary or learn other input methods.

8.2. Character Channel - a Solution to the Compatibility Problem

In this section the Character Channel concept and how it handles the language compatibility in computer systems will be introduced. After a careful analysis of CCDOS and 中文Talk it is clear that the so called Chinese/English compatibility means that an operating system should be able to distinguish two kinds of characters and treat them correctly in all three parts of the system (i.e. input, output, and processing). All software applications supported by the operating system should be able to use both languages.

Based on above observations and analysis of CCDOS and 中文Talk, the author suggests a systematic way to solve the compatibility problem, a Character Channel.

8.2.1.What is a Character Channel

A Character Channel is a means of tracing character-flow in a computer system which can provide important information for multilingual compatibility.

A Character Channel should minimally provide the following information:

- Through which interfaces the characters go?
- How to identify characters of different languages?
- How is the character code changed?

More specifically, a Character Channel should provide:

(1) the information on the character-interface, that is, a character-flow-chart which indicates all interfaces where the characters go through. The interface (called a character-interface) may be an interrupt trap, a switch-byte, a register, etc.

The key issue for a character-flow-chart is that it must indicate which character-interfaces are fixed, that means in any cases that characters must go through those interfaces. When the computer manufacturer (e.g. Apple) releases a new version for the system or third party software developers (e.g. Microsoft) write new software applications for the system, they all follow this rule. It will be discussed later that the concept of a fixed-character-interface will result in a multilingual compatibility.

(2) the information on code-changes:

- the character code set (mainly the internal code).
- the place where code conversion occurs.
- the state of Chinese-flag (e.g. the 8th bit of a character byte, a flag-byte, etc.)

All code-change information is also important for good compatibility.

There may be three kinds of Character Channels

- the Character channel provided by the computer manufacture;
- the Character channel provided by a Chinese software developer;
- the Character channel provided by third party software vendor.

A simple example is shown by Fig. 8.1, the Character Channel for MS-DOS of the IBM-PC.

Character flow chart:

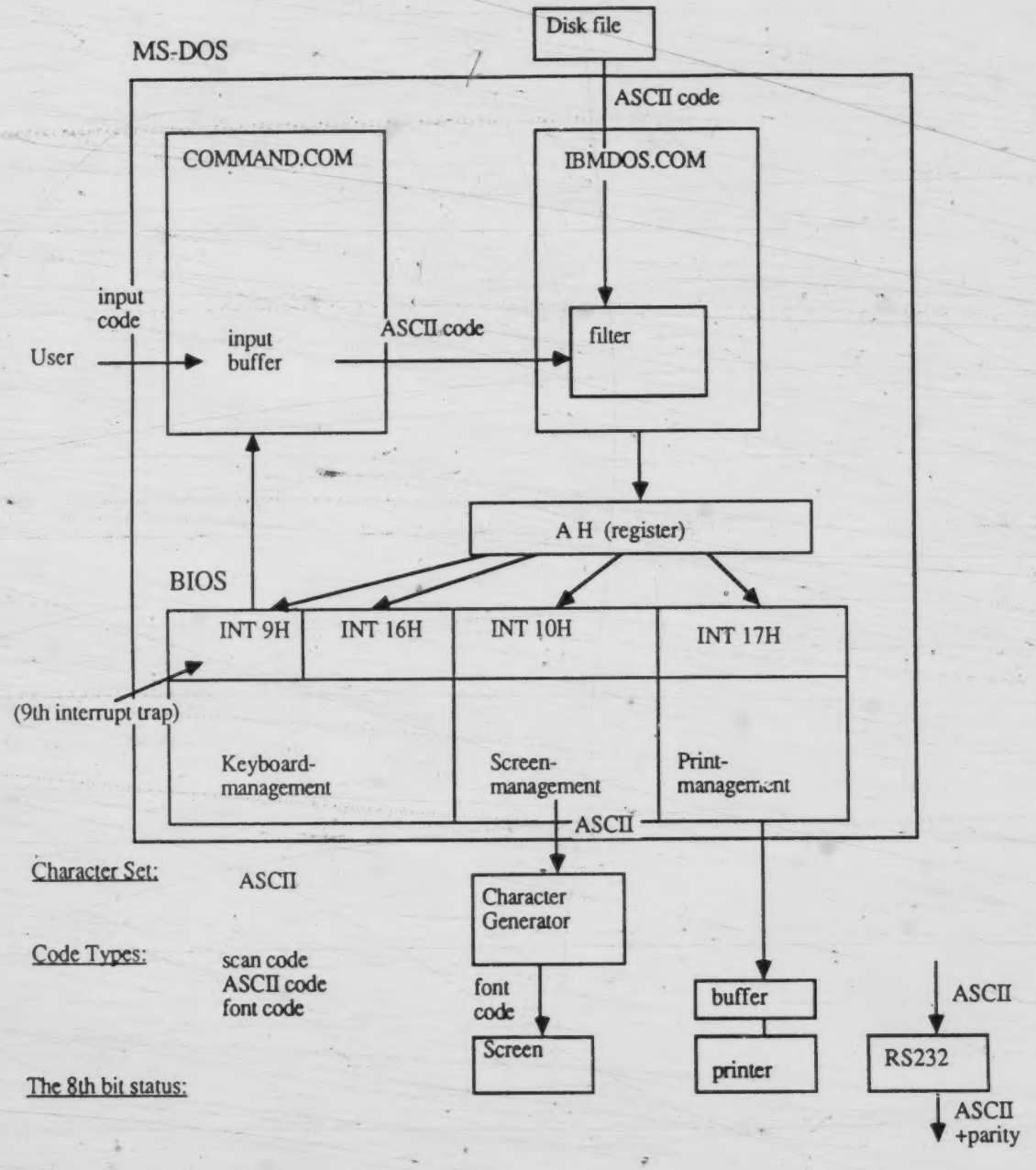


Fig. 8.1 The Character Channel for MS-DOS

8.2.2. Character Channel and Good Compatibility

(1) A Character channel gives computer manufactures, third party developers, and Chinese users a clear idea about the character-flow in a computer system so that they can easily design, modify, and use multi-languages in the system. In other words, by tracing a Character Channel of a computer system, they know where to put a character-filter to distinguish between languages (e.g. Chinese and English), where to insert a new routine or modify original routines to process Chinese and English in different ways, etc.

Note that all character-interfaces are fixed which implies that both Chinese and English characters will be accepted. It is easy to use a filter to distinguish them and treat them by the appropriate routine. A compatible system is thus achieved.

On the other hand, if interfaces are not fixed then compatibility cannot be guaranteed. Assume a character-filter is put at character-interface A, while some applications software may go through interface B (no filter there) then Chinese characters will be treated as something else.

(2) It may help to develop compatible software applications. Namely, a third party software developer may write applications strictly according to the Character Channel of a computer system, that is, restrict character-flow to the fixed character-interfaces of the system, then compatibility will be guaranteed.

(3) It may help to solve the compatibility problem of an original version of an application. Suppose there exists software with some compatible problems. This is very common due to the fact that many applications like MacWrite was developed before 中文Talk was released, therefore, it is very likely that MacWrite does not call any new routines in 中文Talk to support Chines/English compatibility. Based on the Character Channel of the computer system (e.g. 中文Talk) it could not be hard to trace the character-flow of the application (e.g. MacWrite) and find out where it distinguishes Chinese characters and non-Chinese symbols and solves the problem by modifying the related routine or replacing the original routine with a Chinese/English compatible routine.

(4) The code-change information in a Character Channel can help to design good internal code for Chinese characters which is the central issue of designing a compatible system. For example, assume the code set of the original system is ASCII, then the internal Chinese code can be decided

that does not conflict with the ASCII code set. In other words, any Chinese character will not take the internal code which is contained in the ASCII set.

(5) It can guarantee the compatibility for a changing computer system and software applications. When new version of a computer system is released a new Character Channel will be attached to it and so is the new application. In this way users will know exactly which version of a system can work with which application (i.e. the compatibility is guaranteed by this match). If an software is not Chinese/English compatible, then its Character Channel may be checked to find the proper version of the computer system which supports it.

In short, the usefulness of a Character Channel is that it turns a formidable compatibility problem into a much more understandable matter.

The Character Channels for CC-DOS and 中文Talk are given below in order to illustrate the above points.

8.2.3. Developing Character Channels for CC-DOS and 中文Talk

(1) The Character Channel for CC-DOS.

Assume that the IBM Company and Microsoft provided a Character Channel for MS-DOS already (Fig. 8.1), then a Chinese/English compatible operating system may be developed based on the Character Channel with comparatively less effort. The result is another Character Channel, the one for CC-DOS (Fig. 8.2).

Before proceeding, two key issues on Chinese/English compatibility are:

(a) There are three aspects involved in designing a compatible system which are input, processing and output of both languages.

(b) Choosing an internal Chinese code is the central issue for processing-compatibility, because all subsystems (modules) of the Chinese system will be designed based on the internal code.

The steps for developing the Character Channel for CC-DOS are the following:

Step 1. Determine the internal Chinese code.

Fig. 8.1 shows that ASCII code is the internal code for MS-DOS which is a 7-bit code. To avoid conflict between internal Chinese code and ASCII code, a meta-GB code may be used as the internal Chinese code which is the GB code (7-bit Chinese Government Standard code) with its 8th

bit set to "1" to indicate a Chinese character. To represent 6763 Chinese characters of GB, a 2-byte code is chosen.

Step 2. Determine the code conversion.

Fig. 8.1 shows that there is a conversion from input code to ASCII code in the Input module of MS-DOS (i.e. COMMAND.COM), so there should be a conversion from the Chinese input code (e.g. Pinyin code, stroke code, etc.) to the internal Chinese code.

Fig. 8.1 shows that there is a conversion from ASCII to the font representation for screen display, so there should be a conversion from internal Chinese code to the Chinese font code. Actually it is a bit map because of the pictorial nature of Chinese character and the graphics mode is necessary for output Chinese characters. In addition, a large Chinese font library replaces the simple character generator because the character-generator is not capable of generating 6763 Chinese characters.

Fig. 8.1 shows that the ASCII code with its 8th bit as parity-bit is used at the RS232 communication interface, so the two-byte internal code of the Chinese characters has to be changed to a communication code since its Chinese-flag-bit (the 8th bit is set to "1") is used as a parity bit now.

Step 3. Determine the Character interface for Chinese characters.

This is the most important part for Chinese/English compatibility.

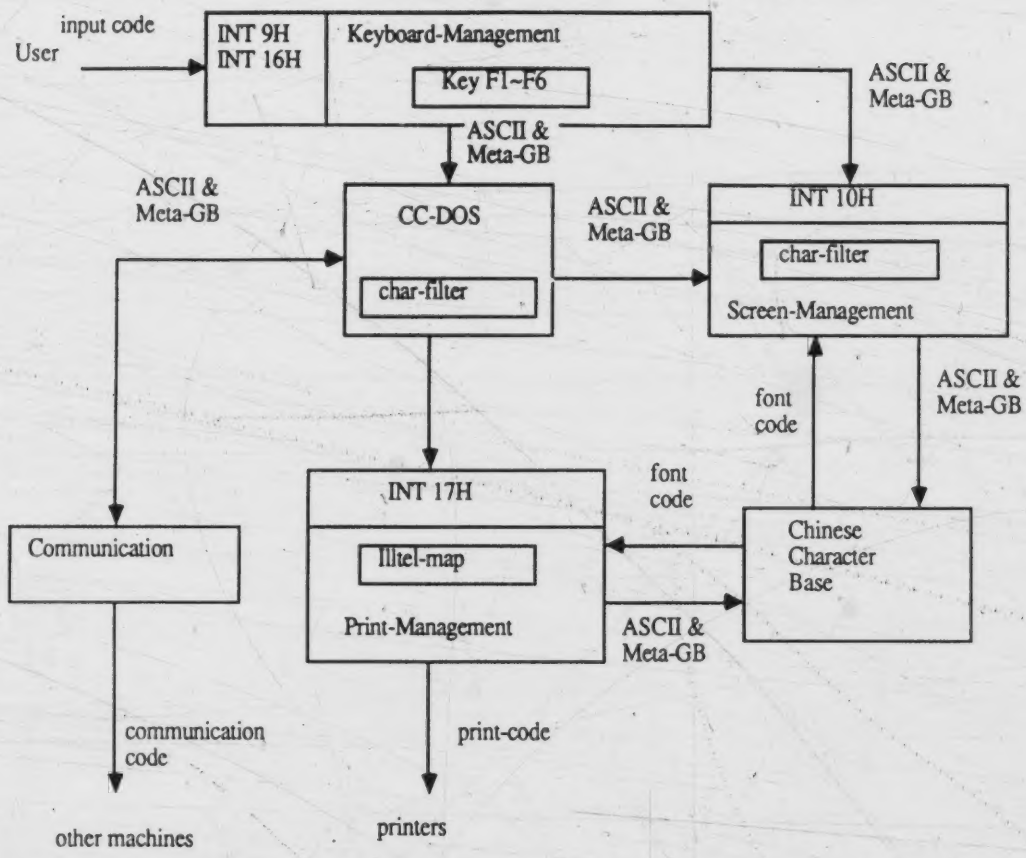
Fig. 8.1 shows all character interfaces where English characters (i.e. ASCII code) will go through. According to the fixed-character-interface rule Chinese characters should use the exact same interfaces. For example, the Chinese characters should enter the system through 16H interrupt trap, should enter screen-management module through 10H interrupt trap, etc. Then character filters for distinguishing Chinese/English characters should follow these character-interfaces so that Chinese and English characters can be treated in different ways (e.g. ASCII code will be displayed as English characters and the two-byte meta GB code will be displayed as Chinese characters on a screen.)

The key point is that no Chinese character will escape checking by character filters because the character-interfaces are fixed with filters following them. In this way compatibility is achieved.

Finally, the Character Channel for CC-DOS is formed (Fig. 8.2). Similarly, the Character Channel for 中文Talk can be built (Fig. 8.3).

Note that the Character Channel for 中文Talk is not accurate because of the complexity of the Macintosh system and the shortage of references for 中文Talk.

Character Flow Chart:



Character Set: ASCII and GB 2312-80

Code Types:
input code
ASCII code
Meta-GB code
font code
communication code (e.g. GB or 4-byte code)

The 8th bit status:
set to "1" to indicate Chinese character
used as parity bit for communication

Fig. 8.2 The Character Channel for CC-DOS

Note: the character-interface and the character-filter are crutial parts

of a Character Channel. In the figure they are sown in the way below

- the Character-Interface is represented as a block attached to a management-module (e.g. INT 10H is attached to the Screen-Management-Module).

- the Character-Filter to separate Chinese and English character is represented by a single small box (e.g. Illtel-map in the Print-Management-Module).

Explanation:

Key F1~F6 -- the function keys for users to switch between ASCII and Chinese characters which are also character-filters to inform the system about ASCII code and Chinese code.

Char-filter -- they are in CC-DOS and Screen-management blocks respectively and have the same function, that is, to separate ASCII and Chinese code (meta-GB code).

Illtel-map -- a switch-byte to distinguish 2-byte-code (meta-GBcode), 4-byte code, and ASCII code.

INT 9H -- the 9th interrupt trap (9H means 9 is a hex-number). so are INT 10H, INT 16H, and INT 17H, that is, they are all interfaces to various management modules which perform character input, output, and processing.

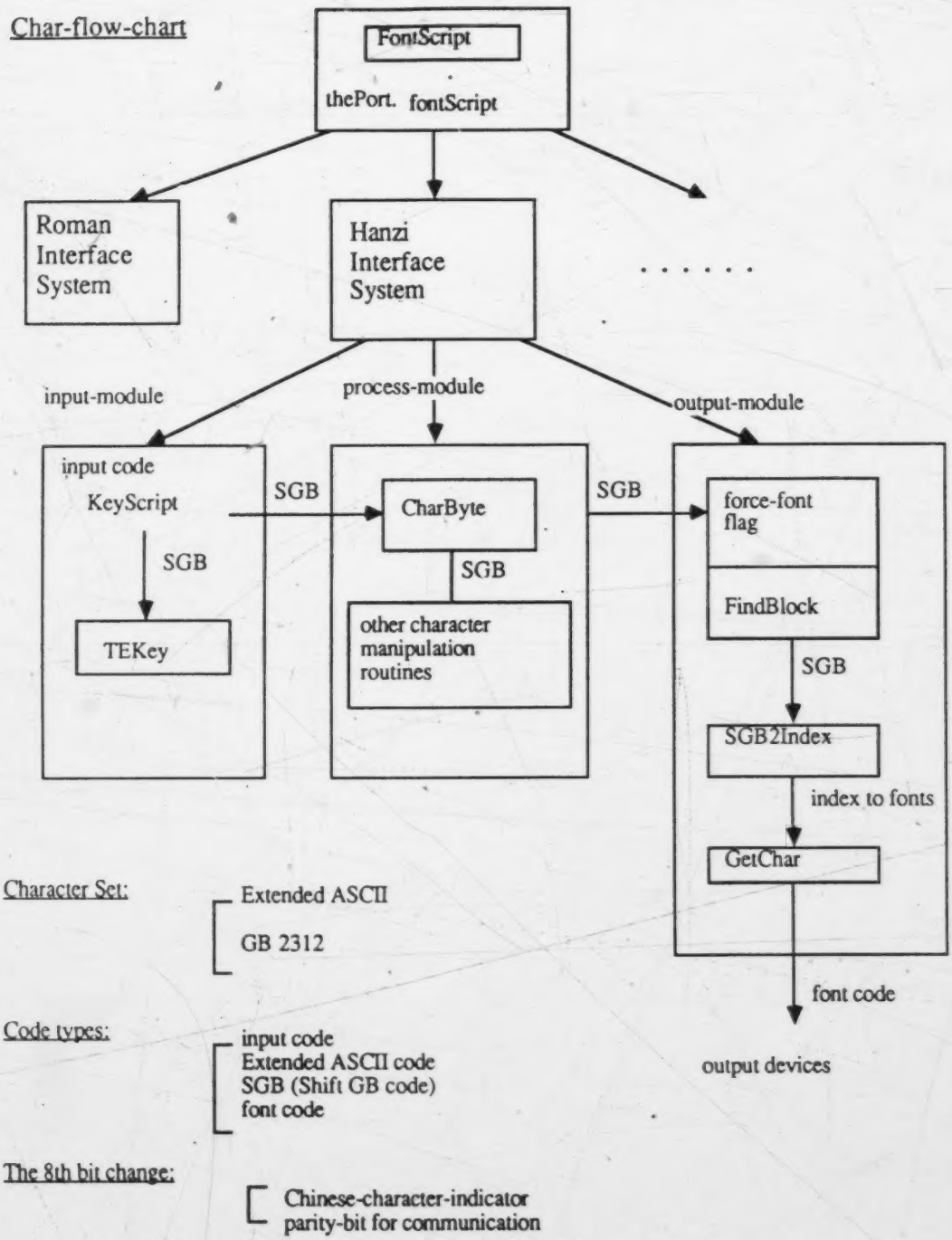


Fig. 8.3 The Character Channel for 中文 Talk.

Brief Explanation:

FontScript -- a routine in Script Manager which checks the font-script field of **thePort** to determine if the script is Roman or Chinese.

KeyScript -- a routine in the Script Manager which determines keyboard script and input methods. It also convers the input code to SGB code.

TEKey -- a routine in HIS which takes a Chinese character (SGB) from the input buffer and separates its first byte and the second byte.

CharByte -- a routine in the Script Manager which takes a SGB code and separates its first byte from the second byte.

force font flag -- a flag in the output module when it is on both ASCII and Chinese fonts can be drawn.

FindBlock -- a HIS Utility routine for determining which characters in a string are Chinese, and which are Roman.

SGB2Index -- a HIS Utility routine for converting from SGB code to the font index.

GetChar -- a routine of the Font Manager in HIS which finds a particular character in a character image file.

8.3. High Quality Character Generation

In Chapters 3 and 6, the technical difficulties faced by Chinese output system are described. They are: huge font storage, fast access, and high quality character generation. Since the cost of hardware is decreasing and optical disk is available now, many believe that in near future, a complete font library will be available to every Chinese users which store a complete Chinese character set with all styles and sizes on an optical disk. The font takes dot-matrix form so that it can be accessed and output very efficiently. With this approach, all complicated on-line generation techniques become meaningless such as data compression, drawing and filling algorithms for skeleton and outline fonts [ZHANG84, GUO85].

However, this approach is questionable due to the following observations:
(1) it is difficult to determine the size of a "complete font library". If it is a font library containing

60,000 characters of the 康熙字典 (Kang Xi Dictionary), then ordinary people will not accept it because 90 % characters of the font library are not useful for them, even if the cost of an optical disk is affordable. If it is a font library relatively larger than GB 2312-80, e.g. 10 to 20 K characters, then in some rare case, a desired character may not be in the library. Therefore, on-line generation is still needed.

(2) Because of the pictorial nature of Chinese characters, it is almost impossible for a font library to include all styles and sizes. Take a popular style, Song, for example, there are many varieties of Song styles such as Imitation Song, Wide Song, Narrow Song, Heading Song, etc. Needless to say the countless numbers of the calligraphic styles.

As a result, on-line generation is still necessary even if optical disk is affordable by ordinary people. Based on the research of the thesis, a better approach will be an integrated font generation scheme, namely, providing users with a relatively "complete" font library and a friendly on-line generation environment. This section will focus on the development of such on-line generation environment.

8.3.1 Problems of the Available On-line Generating Facilities

Currently, most Chinese systems provide on-line character generating facility, however, many of them are neither efficient nor user friendly. The main problem is the deficiency of drawing high quality characters. Most of them display a dot-matrix for users to draw characters on it. Usually, it is a small dot-matrix (e.g. 16*16 for the Macintosh). It is very difficult to draw a character with acceptable quality on such a simple dot-matrix. Even if a large dot-matrix (e.g. 24*24 for KaiHin CWP) is provided, it is still difficult for ordinary people to design a character in good shape, because the structure of a Chinese character is very standard and rigid. A slight change may result in an ugly character.

A friendly on-line generating environment is described below, that enables users to generate high quality characters with arbitrary styles and sizes by using general-purpose hardware and software (i.e. the Macintosh, LaserWriter, Fontographer, and PostScript language).

8.3.2 A Friendly On-line Generating Environment

This approach establishes a stroke library in both geometric language and PostScript. It uses the strokes to construct HQCC in the following way:

Step 1: use Fontographer (a font generating application for the Macintosh) to generate strokes for

a style (e.g. Kai or Li style) [FONT86]. Figure 8.4 shows some strokes in Li style. There are about 10 strokes chosen for each style with about 8 variations for each stroke. Based on the research by the Beijing University, the quality of characters are guaranteed by the strokes (See chapter 6). A font in the Macintosh contains more than two hundred positions, so it is enough for storing all strokes for a style.

Step 2: convert the geometric description of the strokes in Fontographer to a PostScript program. Since the position of control points are indicated in Fontographer, the conversion is straightforward. Note that the PostScript language is a very popular page-description language for laser printing [POST85].

Step 3: use the strokes in the stroke library (built by steps 1 and 2) to compose a character required. Namely, based on the geometric description of strokes in Fontographer (Fig. 8.5), modify their corresponding PostScript program (e.g. translate, scale, and rotate) to compose the character. Note that this process usually takes two steps: first generate a prototype character in PostScript, then modify the program to obtain the exactly the artistic character we want and run the program on the Macintosh LaserWriter to generate the character (Fig. 8.6 to 8.8). The following is a part of a PostScript program that generates characters 巴拉. Note that each stroke is a module in the program and a character is formed by calling the modules. So the main program is very simple.

```
/LA                 % the module for drawing the character "拉"
 {heng1
  shugo
  ti
  heng2             % draw a stroke "—"
  dian1
  dian2
  pie
  heng3
 } def
```

```
gsave
    1 30 translate          % put "拉" in the proper position
    0.6 0.6 scale           % obtain the required size
    LA                      % draw character "拉"
    18 0 translate
    0.4 0.4 scale
    BA                      % draw character "巴"
grestore
```

The main problem for any stroke-based (or radical-based) Chinese character generation is the distortion in quality introduced by composition. However, for the casual user the problem is not serious. That is, a prototype of the character can be provided by the client or other sources (e.g. a Character-Shape dictionary) and is digitized and copied into Fontographer font. Then, PostScript programs for strokes can be modified against the prototype to avoid any distortion. After the PostScript program for the prototype is obtained, it is easy to modify the program to generate a specific artistic character needed by the client.

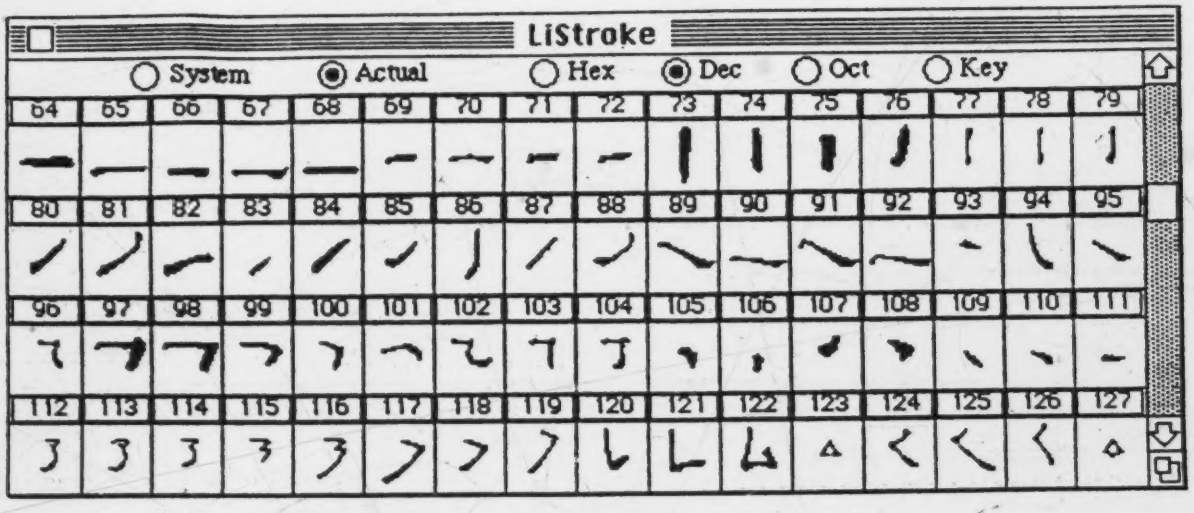


Fig. 8.4. Some Strokes in Li-Style Stroke Library.

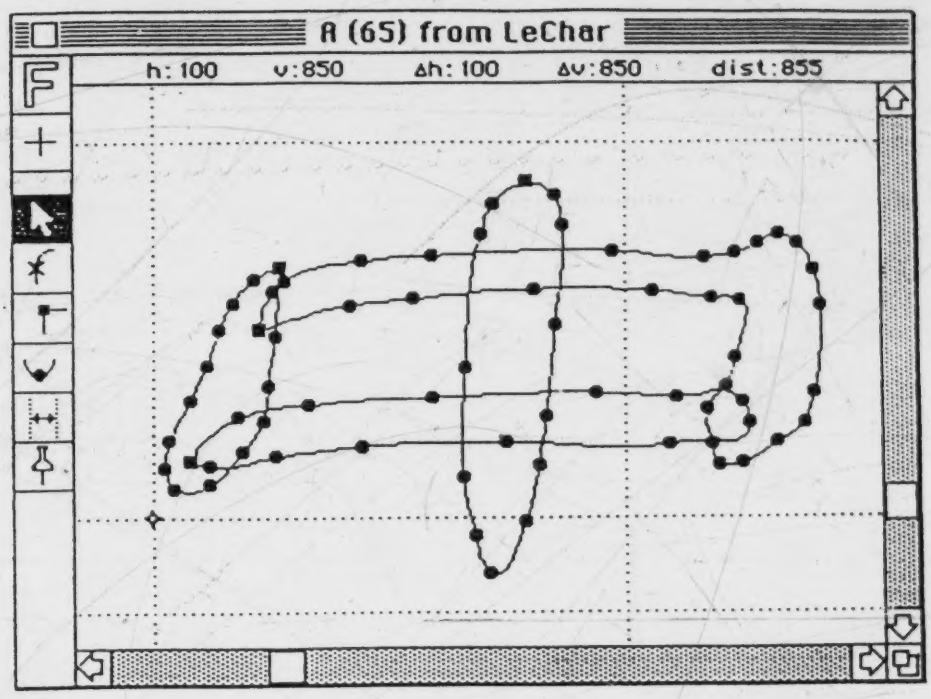


Fig. 8.5. The Character 中 ("center" in English) Formed by Stroke-Composition.

8.3.3 Discussion

Usually, it is not necessary to generate characters by PostScript programming. In fact, only some casual users such as a commercial graphics company need such high quality characters. For ordinary people, the characters generated by Fontographer is good enough. The key point with this approach is that it is a friendly on-line generating environment, because

(1) the prototype makes the task much easier and character quality is guaranteed;

(2) the new characters are kept in a Macintosh font which may be called "User font library" and the font library may be called "Main font library". The former can be built according to different requirements so that users can access the character conveniently. Since the font has one to one mapping with the keyboard, character input can be efficient.

(3) this approach can satisfy any special requirements without restriction. For example, in some cases, a user may need a few artistic characters for a document (note: it is almost impossible for the main font library to keep such characters). Then, the characters can be generated by this approach without any difficulty.

Fig. 8.6. Some Song-Style Characters with Different Sizes and Colors.

Fig.8.7. Variations of Li-Style Characters by Scaling.

Fig.8.8. Some Artistic Characters Based on Li-Style Prototype.

8.4. Conclusion

In this chapter, some recommendations are given: CDPinyin, Character Channel, and High Quality Character Generating. CDPinyin is an improvement to TianMa CWP which is to help users input unknown characters by phonetic method so that they need neither to learn a shape-based method, nor to look up dictionary. The results of experimental tests show the obvious advantage of CDPinyin over looking up dictionary. More improvements are required before Pinyin method becomes a national standard input method.

The concept of Character Channel is introduced that provides information about

code-conversion and character-interfaces where characters will pass through. Character channels for CC-DOS and 中文Talk are illustrated. These examples show that a Character Channel provides a systematic way to discuss and explore the compatibility problem (i.e. make it an understandable problem) and to bring computer manufactures, third party software developers, and Chinese users together to build an integrated Chinese/English compatible system. One point worth mentioning is that, in the past, a Character Channel was not necessary due to the fact that the internal code of a computer system was very simple (e.g. ASCII or EBCDIC code) and it can be treated in simple way in all three phases (i.e. input, processing, and output). Whereas when processing ASCII characters together with Chinese characters (which has a very large ideogram character set), a systematic means like a Character Channel is indispensable. In addition, the concept and method of the Character Channel are not limited to Chinese/English compatibility only and they can be easily applied to other language compatibility problem (e.g. English/Arabic or multi-lingual compatibility). In fact, mono-lingual software is less competitive in today's international market. To enable computer vendors to deal with language barriers in a cost effective way, a global solution must be developed. Multi-lingual compatibility problem is attracting more and more attention of computer scientists [HO88, CEP86].

The objective of High Quality Character Generation is to provide users with a friendly on-line generation environment, which is a compensation to the font library of the Chinese system. The environment supports high quality character on-line generation by using general-purpose hardware and software (i.e. the Macintosh, LaserWriter, Fontographer, and PostScript language). It can generate high quality characters with arbitrary styles and sizes and build a user's own small font library. This approach together with the main font library in the computer system forms an integrated Chinese font system.

Chapter 9
CONCLUSION

This thesis is intended to discuss Chinese computing from the perspective of Chinese system developers and analysts, namely, it investigates such problems as what is the systematic approach to develop an integrated Chinese system, what is the central issue of Chinese computing, etc. The discussion is based on an in-depth survey of four most important areas of Chinese computing: computer representation of Chinese characters, Chinese system software, Chinese word processing systems, and Chinese electronic publishing systems. They reflect the most outstanding features of Chinese computing.

9.1. Conclusion

Systematic approach has obvious advantage over Patching approach. The former treats Chinese processing as an integrated part of computer system in order to obtain the best compatibility. According to this approach, some standard such as Character Channel should be established and followed by three parties: computer manufacturers, third party software developers, and users. Also, a Chinese system should be developed on a solid foundation (i.e. hardware and operating system) such as graphics-based computers.

Graphics-based computers provide an environment powerful enough to solve the bottleneck of Chinese computing, Chinese input and output, because many facilities required by Chinese computing are available in graphics-based computer systems already, while with text-based computers a Chinese system developer has to design the facilities from scratch.

The central issue for obtaining an integrated Chinese system is how to obtain Chinese/English compatibility in all three parts of a Chinese system, i.e. input, editing, and output of Chinese. The compatibility problem is so important that all aspects of Chinese computing and all applications (e.g. word processing, database, network, etc.) are affected by it. Graphics-based computers such as the Macintosh provide a good foundation for developing an integrated Chinese system with very good compatibility. A concept of Character Channel is given that tries to solve the compatibility problem in a systematic way. A Character Channel is a standard for restricting character-flow in Chinese systems, namely, character flows between different modules in a computer system by fixed interfaces. The interfaces are fixed by computer manufactures so that software developers can follow this rule to design a Chinese/English compatible application. Computer users use the system and software applications based on the understanding of the rule (Character Channel).

Although Chinese word processing (CWP) has not been given enough attention in China, it deserves the most important application in Chinese computing due to the fact: CWP reflects two most outstanding features of Chinese computing, i.e. input and output of Chinese characters. The very detailed criteria for evaluating CWP are established in order to evaluate existing CWPs and design a new CWP. The available CWPs, i.e. TianMa, Kaihin, and ZhongWen Talk are evaluated according to the criteria. The evaluation results are compared with experimental tests. The main idea is to evaluated a CWP in its entirety, namely, whether the user-CWP interface is friendly, whether the CWP provides a good foundation so that users can customize the CWP for their requirements, etc.

Chinese electronic publishing (CEP) is another important feature of Chinese computing, the output of Chinese. By analyzing the Hua Guang CEP developed by Beijing University, it is clear that Chinese output is as difficult as Chinese input. To obtain output efficiency and quality with an acceptable storage overhead, various techniques are invoked such as run-length codes, radical methods, and stroke methods. Laser printing is also playing an important role in this area. The experiment of HQCC Generating (in chapter 8) shows the feasibility of generating high quality characters by using popular laser printing facilities. More importantly, it provides a friendly on-line generating environment as a compensation to the main font library. This approach together with the main font library forms an integrated Chinese font system.

Many other research areas of Chinese computing are not covered in the paper either because they do not show the unique features of Chinese computing (e.g. Chinese computer networks) or because there is a shortage of reference materials (e.g. Chinese computer architectures).

9.2. Further Work

This thesis reveals many interesting and significant research topics. Some possible projects are:

(1) To further develop the concept of Character Channel, namely, to analyze an application software and the relation between the application and the operating system supporting it (e.g. MacWrite and 中文Talk of the Macintosh). To investigate the feasibility of setting up a Character Channel standard for computer manufacturers and application software developers.

(2) To continue the study of Pinyin input method, namely, to find a way to solve the accent-tolerant and error-tolerant problems, to implement Character Decomposition Pinyin (possibly with the cooperation of TianMa CWP developers).

(3) To implement a simple CWP on the Macintosh and use it as a test-bed for experiments of an integrated Chinese computing system.

(4) To thoroughly analyze the relationship between a graphics-based computer and Chinese processing, and based on the analysis to design a Chinese computer architecture or implement a Chinese system for a graphics-based computer such as the Sun workstation.

References

[CEP86] Rodger E. Burkley, Introducing Chinese VMS Software (CEP) and The VT82 - Digital's Hanzi Character Products, Far East Software Services of DEC, Mar. 1986.

[CHAN88] Y.S. Chan et al. Chinese Character Generator, *Proceedings of 1988 International Conference on Chinese Computing*, pp. 325-328, 1988.

[CHIN87] A.Chin and F.Wong, An Intelligent, Knowledge-based Chinese Input System, *Computer Processing of Chinese and Oriental Languages*, Vol. 3, No. 1, pp. 25-32, May, 1987.

[CUNIX89] Universal UNIX Operating System, People's Daily (oversea edition), P.R. China, March 15, 1989.

[DICT80] The Modern Chinese Dictionary, Commercial Press, P.R. China, 1980.

[FONT86] Fontographer User's Guide, AltSys Co. 1986.

[FRED84] Fredirick Holtz, Using and Programming the Macintosh, TAB Books Inc. 1984.

[GB81] Code Table of Graphic Characters, GB2312-80, 1981.

[GE87] Zhang Ge, The Users Manual of Chinese Macintosh System, TianJin, P.R. China, 1987.

[GUO85] Guo Pingxin and Zhang Songzhi, (editors), The Technology of Hanzi Information Processing, Defense Industry Publishing Co. Dec. 1985.

[GUO86] Guo Yefang, A Summary of the Techniques of Input Chinese with Keyboard, The Technical Research Institute of XinHua News Agency, Nov. 1986.

[GW84] The Chinese System for IBM-PC and its Compatible GreatWall-100 microcomputer, the Research Institute of Electronic Technology, the Ministry of Electronic Industry, China, 1984.

[HIS87] Mark Davis, et al. Hanzi Interface System ERS, June, 1987.

[HO88] Y.F. Ho, et al. A Multi-lingual Information System and its Design Considerations, *Proceedings of 1988 International Conference on Chinese Computing*, pp. 595-598, 1988.

[HUANG86] S.J.Huang, et al. Chinese-English Compatible Microcomputer Systems, *Proceedings of 1986 International Conference on Chinese Computing*, pp. 344-348, 1986.

[INSV88] Inside Macintosh Vol.V, 1988.

[JIANG85] Jiang Yuxing, A Discussion of the Standardization of Input Methods for Chinese Computers, *Chinese Information Processing* (in Chinese), pp.10-12, Feb. 1985.

[KAIHIN86] BrushWriter, Chinese Word Processing Software, Kaihin Research Inc. 1986.

[LO85] Lo S.Y., A Scientific Model for Comparing Various Methods of Inputting Chinese Characters into Computers, *Computer Processing of Chinese and Oriental Languages*, Vol. 2, No. 1, pp. 36-58, May, 1985.

[MOON88] Y.S. Moon and Y.P. Szeto, Efficient Construction of High Quality Chinese Font Libraries, *Proceedings of 1988 International Conference on Chinese Computing*, pp. 262-265, 1988.

[MORG85] C.L. Morgan, Hidden Powers of the Macintosh, The Waite Group Inc. 1985.

[NEWS87] Hanzi Frequency of Usage for News Infarmation, Xin Hua Communication, Jan. 1987.

[NI85] Ni Guangnan, The User Interface of A Chinese PC (LX-80), *Proceedings of Compint85*, pp. 771-773, Sep. 1985.

[PD88] People's Daily (oversea edition), P.R. China, June 1988.

[POST85] PostScript Language Reference Manual, Adobe Systems Inc, 1985.

[QIAN86] Qian Peide, The Analysis of CC-BIOS, Su Zhou University Press, P.R. China, 1986.

[SUN86] Sun-3 Architecture: A Sun Technical Report, Sun Microsystem, August 1986.

[TIAN86] TianMa Handbook, International Geosystems Corporation, Vancouvor, Canada, 1986.

[WARD83] D.B. Ward, X. Wang, W.A. Davis, Interactive Creation of Chinese Characters, *Proceedings of Graphics Interface '83*, Edmonton, Canada, May, 1983.

[WEI85] Wei Cui, Evaluation of Chinese Character Keyboards, *Computer Journal*, pp. 54-59, Jan. 1985.

[YU88] Wellington C. P. Yu, A Historical Advancement of the Chinese Language Computing, *Proceedings of 1988 International Conference on Chinese Computing*, pp. 9-21, 1988.

[YUZ85] Yu Zehe, Chinese Computer and Chinese Character Reformation, *Chinese Information Processing* (in Chinese), pp.51-53, Feb. 1985.

[ZHANG84] Zhang Shouxuan, Computer Processing of Chinese Information, National Defence Press, P.R. China, 1984.

Appendix A

Evaluating the Available CWPs by Three Methods

In this appendix, three methods are used to evaluate three CWPs, namely, KaiHin CWP, 中文 Talk, and TianMa CWP. These methods are the Wei's criteria [WEI85], Lo's criteria [LO85] and the proposed criteria. The results are compared to show the validity of the proposed criteria.

A.1. The Results by the Proposed Criteria

(1) Input Method Criteria (Table III):

Note that since TianMa provides only the Pinyin input method, Pinyin is evaluated for all three CWPs.

KaiHin CWP:

Criterion	Level	Points	Explanation
1	C	5	conversion affects thinking due to homonyms indication
2	A	7	most users know Pinyin before using CWP
3	A	2	duplicate code rate is zero and the number of strokes is 3.7 (Pinyin is like English).
4	C	2	4 methods exist
5	C	4	Pinyin is semi-touch-typing and KaiHin's own component method is touch-typing.
6	B	3	needs a printer_controller (note: new version has eliminated it)
7	B	3.75	WIMP (window, icon, mice, and pointer)
Total:		26.75	

中文 Talk:

Criterion	Level	Points	Explanation
1	C+	5.83	conversion affects thinking but better than KaiHin CWP (homonyms are sorted by usage frequency)
2	A	7	most users know Pinyin before using CWP
3	A	2	duplicate code rate is zero and the number of strokes is 3.7 (Pinyin is just like English).
4	C	2	4 methods exist
5	C	4	Pinyin is semi-touch-typing and KaiHin's own component method is touch-typing.
6	A	4	no extra hardware needed
7	C	2.5	worse than KaiHin, it's not a CWP
Total:		27.33	

TianMa CWP:

Criterion	Level	Points	Explanation
1	B	7.5	easy and smooth typing with some errors which need manual-correction
2	A	7	most users know Pinyin before using CWP
3	A	2	duplicate code rate is zero and the number of strokes is 3.7 (Pinyin is like English).
4	D	1	one method (Pinyin)
5	B	6	about 50 Hanzi/min, touch-typing
6	B	3	need TianMa ROM card
7	A	5	phrase input, automatic conversion
Total:		31.5	

Result: TianMa is the best mainly because of its phrase input, smooth typing, and automatic conversion.

(2) Pinyin Input Criteria (Table IV):

KaiHin CWP:

Criterion	Level	Points	Explanation
1	D	2.5	stop to indicate homonym and hit rate is low
2	D	1	sorted according to GB 2312-80 only
3	D	0.75	exact one way of Pinyin
4	D	0.75	no tone
5	D	1.75	no phrase base
6	D	1.75	standard Pinyin only
7	D	0.75	look up dictionary
8	D	0.75	no way to help user
Total:		10	

中文 Talk:

Criterion	Level	Points	Explanation
1	C	5	hit rate is high
2	C	2	homonym is sorted by usage-frequency
3	C	1.5	partial-Pinyin tolerant
4	A	3	allowes both with and without tones
5	D	1.75	no phrase base
6	D+	2.45	with some intelligence although it is standard Pinyin
7	D	0.75	look up dictionary
8	D+	1.05	allow partial-Pinyin input
Total:		17.5	

TianMa CWP:

Criterion	Level	Points	Explanation
1	A	10	typing without need of indicating homonym
2	A	4	automatic indication
3	C	1.5	phrase input
4	C	1.5	user feels no inconvenience
5	A	7	intelligent Hanzi and phrase base
6	D	1.75	standard Pinyin only
7	D	0.75	look up dictionary
8	C	1.5	can modify Hanzi/phrase base
Total:		28	

Result: TianMa is the best mainly because of its phrase input and automatic conversion.

(3) Output Criteria (Table V):

KaiHin CWP:

Criterion	Level	Points	Explanation
1	D+	2.33	with many artistic character styles and only one calligraphy-style, Song style.
2	B	2.25	more than 5 character sizes provided
3	C	1	in graphics mode (because of the Macintosh)
4	A-	2.75	need to insert special symbols to mark the style
5	A	2	capable of modifying characters
6	C	5	insert special symbols
7	C+	4.67	worse than 中文 Talk
8	B	3.75	laser printer
Total:		23.75	

中文 Talk

Criterion	Level	Points	Explanation
1	D+	2.33	with many artistic character styles and only one calligraphy_style, Song.
2	A-	2.75	better than KaiHin if MacPaint used
3	C	1	graphics mode (because of the Macintosh)
4	A	3	changeable at all levels (MacPaint)
5	A	2	capable of modifying characters
6	A	10	WYSIWYG
7	B-	5.33	better than KaiHin
8	B	3.75	laser printer
Total:		30.16	

TianMa CWP:

Criterion	Level	Points	Explanation
1	C	3.5	with both simplified and classical styles and ImitativeSong style only
2	C	1.5	only 2 sizes
3	C	1	draw manually
4	C	1.5	characters within a paragraph are unchangeable
5	B	1.5	a problem exists with right margin for big characters
6	B	7.5	both simplified and classical style shown on screen
7	C	4	slow display and printing
8	D	1.25	ordinary devices (dot matrix printer)
Total:		21.75	

Result: 中文 Talk is the best in Output aspect mainly because of its high quality output devices and the Chinese operating system (higher efficiency than KaiHin).

(4) Editing/Formatting Criteria (Table VIII):

KaiHin CWP:

Criterion	Level	Points	Explanation
1	B	3.75	WIMP (window, icon, mice, and pointer)
2	D	0.625	none
3	B	1.875	general page Formatting
4	D+	0.833	provides a few function (e.g. tab)
5	D	0.625	a pure text processor
6	B-	3.33	on-line help
Total:		11.04	

中文 Talk:

中文 Talk is not a true CWP and it does not have editing/formatting functions.

TianMa CWP:

Criterion	Level	Points	Explanation
1	D+	1.67	no window system, but menu provided
2	B	1.875	no footnotes
3	B	1.875	general page Formatting
4	B	1.875	no superscript/subscript
5	D	0.625	a pure text processor
6	B-	3.33	on-line help
Total:		11.25	

Result: KaiHin is slightly better than TianMa CWP because of its window system.

FINAL RESULT:

	KaiHin	中文Talk	TainMa
Pinyin	10	17.33	28
Editing	11.04	---	11.25
Output	23.75	30.16	21.75
Total (with Editing)	44.79	47.49	61
Total (without Editing)	33.75	47.49	49.75

By comparing the total score of three CWP, TianMa is the best among them. 中文Talk, however, provides a better foundation for CWP development (based on the result of Output Criteria).

A.2. The Results by Applying Wei's Criteria

Applying Wei's evaluation criteria described in [WEI85], the results for the CWP are the following:

No. Criteria	KaiHin	中文 Talk	TianMa
1 Recognition of coding symbols	10	10	10
2 Pronunciation	10	10	2
3 Coding	8	10	10
4 Ratio of coding to key symbols	5	4	6
5 Single-level coding	5	5	5
6 Coding rule consistency	5	5	5
7 Keyboard size	10	10	10
8 Touch-typing	1	2	3
9 Number of keystrokes	7	7	7
10 Duplicate codes	10	10	10
11 Vocabulary	3	3	3
12 Readable codes	3	3	3
13 Costs	12	12	12
14 Compactness	2	2	2
Total:	91	93	87.5

Comment: TianMa turns out to be the worst CWP by Wei's criteria, which is contrary to the results of the experimental test (see section 5.5).

A.3. The Results by Applying Lo's Criteria

Applying Lo's evaluation criteria described in [LO85], the results for the CWP are the following:

Criteria	English	KaiHin	中文 Talk	TianMa
1 Keyboard size				
Beginner(L1)	1	1	1	1
Expert(L2)	1	1	1	1
2 Symbols (S)	1	1	1	1
3 Analysis steps in typing(P)	1	1.65	1.65	1.65
4 Degree of difficulty in typing(V)	1	0.64	0.64	0.64
5 Indexing system for data base				
Memory effort (I1)	1	1	1	1
Searching effort (I2)	1	0.64	0.64	0.64
total:	7	6.93	6.93	6.93

Explanation:

(1) Keyboard size: all three CWP use standard English keyboard.

(2) The number of symbols a key represents: typing Pinyin is the same as typing English.

(3) Analysis steps in typing: which means the conversion from Chinese character to sound (or component).

According to Lo's formula

P = 23.6 for English

P = 38.79 for Pinyin which is 1.65 (normalization)

(4) Degree of difficulty in typing: it is related to the speed of typing.
Pinyin has less keystrokes than English so it should be quicker than English typing.

(5) Indexing system for data base: in which the Memory effort (I1) stands for the effort one needs to remember the indexing (i.e. ordering).

Since both Pinyin and English consist of 26 Latin scripts, they have the same I1. Searching effort (I2) measures the work required to look up a word in an indexing system. Since Pinyin code is shorter than an English word (3.7 : 5) I2 for Pinyin is 0.64.

Comment: All CWPs obtain the same score, that is, Lo's criteria is not capable of distinguishing between them.

A.4. Comment on the Results of Applying Three Methods

It has been shown that the proposed evaluation criteria can distinguish the main differences between three CWP and obtain a proper result, whereas Wei's and Lo's criteria is not capable of distinguishing between the CWP because of the following reasons:

(1) Wei's and Lo's methods are for evaluating Input method (or keyboard) instead of the whole Chinese word processor system.

(2) Wei's and Lo's methods do not distinguish between different Pinyin input methods but they treat the standard Pinyin as the only one.

(3) Wei's and Lo's methods treat the input method as the sum of several separate elements and analyze the elements one by one instead of treating it as the whole (e.g. they do not examine if the input method allows smooth typing or if it provides a friendly interface so that an inexperienced user may interact with it easily).